Hochschule Hannover
Fakultät III – Medien, Information und Design
Abteilung Information und Kommunikation

Möglichkeiten zur Unterstützung der automatischen Spracherkennung in wissenschaftlichen Videos mit Hilfe von Fachterminologie

Masterarbeit im Masterstudiengang
Informations- und Wissensmanagement

vorgelegt von

Letitia-Venetia Mölck
Matrikel-Nr. 1148170
Wintersemester 2012/2013

Erstgutachter: Prof. Dr. Christian Wartena

Zweitgutachter: V.-Prof. Dr. Jutta Bertram

Hannover, den 01.03.2013

Zusammenfassung

Automatische Spracherkennungssysteme (Automatic Speech Recognition - ASR) können derzeit nicht alle Wörter korrekt erkennen und daher noch keine guten Transkriptionen erstellen. Die Qualität der automatischen Spracherkennung wird von vielen Faktoren beeinflusst. Einer davon ist das Vokabular. Je vielfältiger und komplexer die Themen, desto größer die Anzahl der fachspezifischen Wörter ist, die erkannt werden müssen, desto schwieriger ist die Erkennungsaufgabe und desto schlechter sind die Transkriptionsergebnisse. Die Sprachmodelle von automatischen Spracherkennungssystemen müssen durch Training angepasst werden, damit sie auch bei Gebieten mit speziellem Vokabular gute Resultate erzielen können.

In dieser Arbeit wird untersucht, ob der prozentuale Anteil der korrekt erkannten Wörter durch Training des Sprachmodells der automatischen Spracherkennung mit fachspezifischer Terminologie wirksam gesteigert werden kann. Anhand von Ergebnissen der durchgeführten Experimente wird dargelegt, welche Anzahl und Art von Daten benötigt wird, um den Prozentsatz der falsch erkannten Wörter zu senken. Die Ergebnisse der Domänen-Adaption bilden die Basis für den anschließenden Vergleich des fachspezifischen Vokabulars in Vorlesungsvideos und wissenschaftlichen Publikationen, um die Unterschiede hinsichtlich der verwendeten Fachsprache aufzuzeigen. Grundlage und Ausgangspunkt für die gesamte Untersuchung stellt die Erkennung der Fachterminologie und ihre Unterscheidung von der Allgemeinsprache dar.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AM	Acoustical Model
ANNIE	A Nearly-New Information Extraction System
ASR	Automatic Speech Recognition
BASE	British Academic Spoken English
CC BY-NC-SA 3.0 U.S.	Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States
CC BY-NC-ND 3.0	Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported
COCA	Corpus of Contemporary American English
CREOLE	Collection of REusable Objects for Language Engineering
DBIS	Datenbank-Infosystem
DF	Dokumentfrequenz
EML	European Media Laboratory
ETDE	Energy Technology Data Exchange
FIZ	Forschungsinformationszentrum
GATE	General Architecture for Text Engineering
HQ	Häufigkeitsquotient
HTML	Hypertext Markup Language
IDF	Inverse document frequency
IEE	Institution of Electrical Engineers
JAPE	Java Annotation Patterns Engine
KNM	Kompetenzzentrum für nicht-textuelle Materialien
LM	Language Model
MAP	Maximum a posteriori
MIT	Massachusetts Institute of Technology
NE	Named Entity
OAI	Open Archives Initiative
OOV	Out of Vocabulary
PDF	Portable Document Format
POS	Part of Speech
PP	Perplexität (Perplexity)
RTF	Real Time Factor
SQL	Structural Query Language
TF	Term frequency
TIB	Technische Informationsbibliothek
TRANSKRIPTVORL	Korpus mit Transkriptionen von Vorlesungen
WCR	Word Correct Rate
WER	Word Error Rate
WISSPUBL	Korpus mit wissenschaftlichen Publikationen

1. Einleitung

Kennzeichnend für das letzte Jahrzehnt ist die rasante Entwicklung von Informations- und Kommunikationstechnologien. IT-Systeme und Softwarelösungen sind aus dem professionellen sowie privaten Umfeld nicht mehr wegzudenken. In Folge der sich stets weiterentwickelnden Informationstechnik können digitalisierte Informationen und Waren mit hoher Geschwindigkeit erzeugt, angezeigt, bearbeitet, kopiert und übertragen werden. Meinel und Sack (2009, S. 4) sprechen in diesem Zusammenhang von einer Informationsgesellschaft, in der über der realen Ebene eine digitale, virtuelle Ebene entsteht, die ausschließlich aus „digitalen Gütern" besteht. Durch Digitalisierung können alle konventionellen Medientypen wie Text, Musik, Film, Fernseh- und Radiobeiträge zu den digitalen Gütern gehören.

Viele Medienprodukte entstehen vermehrt nur noch in digitaler Form. Bei der Darstellung und Vermittlung von Informationen und Wissen können mehrere Medientypen sinnvoll kombiniert werden. Die so entstandenen digitalen Werke werden als Multimedia bezeichnet (vgl. Meinel und Sack, 2009, S. 12). Für diese digitalen Medien steht eine Vielzahl von Datenformaten zur Verfügung. Für den Vertrieb über unterschiedliche Computernetzwerke müssen die digitalen Medien entsprechend zum Medientyp kodiert und komprimiert werden. Zur Bereitstellung und zum Vertrieb von digitalen multimedialen Produkten wird eine spezifische leistungsfähige Infrastruktur benötigt, die den Anforderungen für Multimedia entspricht.

Digitale Multimedia-Bibliotheken und audiovisuelle Portale verfügen über die adäquate Infrastruktur sowie über das technische Know-how, um die digitalen multimedialen Werke zu sammeln, zu organisieren und zu kategorisieren sowie die passenden Werkzeuge zum Finden der Informationen zu entwickeln und bereitzustellen. Damit die multimedialen Objekte auffindbar sind und genutzt werden können, müssen die im Audio- und Videocontent enthaltenen Informationen mit geeigneten Techniken analysiert, anschließend die Metadaten extrahiert und mittels sprachlicher Repräsentationen bereitgestellt werden. Dies ist von besonderer Bedeutung, da nur dadurch eine Full-Content-Suche nach den in den Multimedia-Inhalten versteckten Informationen ermöglicht wird (vgl. Hauptmann,

Jin & Wactlar, 2000, S. 7). Am Beispiel der Informedia Digital Video Library[1] geben Hauptmann et al. (2000, S. 7-10) einen guten Überblick zum Themengebiet der Content-Analyse von digitalen Videos.

Für die vorliegende Arbeit stellen die Audioanalyse mittels Spracherkennung und die automatische Erstellung von Transkriptionen sowie die Extraktion von Informationen aus der gesprochenen Sprache den theoretischen Hintergrund dar. Es werden die Möglichkeiten zur Unterstützung der automatischen Spracherkennung durch die Erweiterung des Vokabulars mit Terminologie für die fachspezifische Adaption des Sprachmodells aufgezeigt.

Nach Jurafsky und Martin (2009, S. 35-38) ist die automatische Spracherkennung (*Automatic Speech Recognition - ASR*) eine wichtige Komponente von Programmen, mit denen natürliche Sprache verarbeitet wird. Die Aufgabe der Spracherkennung erfordert ein komplexes Wissen aus den Gebieten der Phonetik, Phonologie, Morphologie und Syntax. Bemerkenswert ist, dass relativ wenige Theorien und Modelle aus dem Bereich der Mathematik, Informatik und Linguistik verwendet werden, um das Wissen aus den genannten Gebieten der Sprachwissenschaft zu extrahieren und für die automatische Spracherkennung zu nutzen.

Die zwei Komponenten der automatischen Spracherkennung – das Sprach- und das akustische Modell – werden von einigen Faktoren beeinflusst. Jurafsky und Martin (2009, S. 320-321) betonen, dass Vielfalt und Veränderungen die gravierendsten Aspekte für eine gute Spracherkennung sind und nennen folgende entscheidende Parameter:

- Größe des Vokabulars: Je größer die Anzahl der Wörter ist, die erkannt werden müssen, desto schwieriger die Erkennungsaufgabe und desto schlechter die Transkriptionsergebnisse.
- Flüssige, natürliche Sprache in Form von spontaner Rede oder Gesprächen: Werden die Wörter einzeln, langsam und deutlich ausgesprochen (*isolated words)* ist die Erkennung einfacher als bei flüssiger, kontinuierlicher Sprache.

[1] http://www.informedia.cs.cmu.edu/

In diesem Zusammenhang nennen Stouten, Duchateau, Martens und Wambacq (2006, 1592–1593) drei Arten von Unterbrechungen, die gerade bei spontanem Reden zur Beeinträchtigung der Spracherkennung führen: gefüllte Pausen, Wortwiederholungen und der Neuanfang eines unterbrochenen Satzes.

- Umweltgeräusche und die Aufnahmegeräte.
- Sprecherspezifische Besonderheiten wie Akzent, Dialekt oder eine Fremdsprache, von Nicht-Muttersprachlern gesprochen.

Je mehr Einflussfaktoren aufeinandertreffen, desto größer ist der Prozentsatz der falsch erkannten Wörter (*Word Error Rate* - *WER*).

Der erstgenannte Einflussfaktor stellt den Ausgangspunkt für das Thema der Arbeit dar. Es wird untersucht, ob und inwiefern sich der prozentuale Anteil der korrekt erkannten Wörter steigern lässt, indem man im Web verfügbare Quellen nutzt, um das vorhandene Vokabular einer Spracherkennungssoftware zu erweitern, es mit Fachterminologie anzureichern und so das Sprachmodell zu trainieren.

1.1 Stand der Forschung

Die Anpassung des Sprachmodells (*Language Model Adaptation*) ist nach Jurafsky und Martin (2009, S. 146) einer der meist untersuchten Bereiche. Dies ist immer dann notwendig, wenn das Programm über eine relativ kleine Anzahl von fachspezifischen Trainingsdaten verfügt und eine große Menge Daten aus vielen verschiedenen Bereichen erkennen muss. Das Basis-Sprachmodell kann man trainieren, indem man den Spracherkenner mit einem großen Korpus von Daten trainiert und die neuen Sprachmodelle an das Basis-Modell anpasst. Eine geeignete Quelle von Daten für diese Art von Training ist das Web, das mittels Suchmaschinen nach passenden Begriffen durchsucht werden kann.

Die Ergebnisse der Experimente von Zhu und Rosenfeld (2001, S. 533–536) haben gezeigt, dass die Wortfehlerrate verbessert werden konnte, indem das Web mittels Suchmaschinen nach *n*-gram als exakte Phrasen durchsucht wurde. Die Ergebnisse, die entweder als Anzahl des Vorkommens der Phrase oder als Anzahl von Webseiten erhalten wurden, wurden für die Anpassung des Sprachmodells

verwendet. Die Autoren empfehlen, nur host-spezifische Webseiten abzufragen, um fachspezifische Daten zu erhalten.

Bulyko, Stolcke und Ostendorf (2003, S. 7-9) plädieren auch für die Gewinnung von Trainingsdaten aus dem Web zur Anpassung des Sprach- und akustischen Modells. Diese Daten sollten so gefiltert werden, dass sie zum Stil und zu den Themen der Zielaufgaben der Spracherkennung passen. Besondere Berücksichtigung findet hier die automatische Spracherkennung von Gesprächen. Das geeignete Trainingsmaterial dafür sind Transkriptionen von Sprachaufzeichnungen. Zu diesem Zweck wurde das Web nach chat-ähnlichem Textmaterial (z.B. Forenbeiträgen) durchsucht, da diese Diskussionen im realen Leben ähnlich sind.

Die Adaption der beiden Komponenten der Spracherkennung (Sprach- und akustisches Modell) zugleich mit Hilfe von vorhandenen Korpora war auch Forschungsgegenstand von Lamel, Bilinski, Adda, Schwenk und Gauvain (2006, S. 457-468). Für das Training des akustischen Modells wurden Transkriptionen von Sprachaufzeichnungen verwendet. Zum Training des Sprachmodells wurden Textdaten aus vielfältigen, online verfügbaren Konferenzschriften extrahiert.

Ein ähnlicher Ansatz ist bei Glass, Hazen, Cyphers, Malioutov, Huynh und Barzilay (2007, S. 2392–2395) zu finden. Die Autoren beschreiben die Zwei-Schritte-Vorgehensweise zur Adaption des Sprachmodells der Spracherkennungssoftware innerhalb des Massachusetts Institute of Technology[2] Spoken Lecture Processing Projects. Zuerst wurde das vorhandene themenunabhängige Vokabular erweitert, indem aus unterschiedlichen textuellen Materialien fachspezifische Begriffe passend zu den Zielvorlesungen extrahiert wurden. Die neu erschaffenen, themenspezifischen Modelle wurden mit dem vorhandenen, themenunabhängigen Sprachmodell gemischt, um dadurch eine Adaption des Basismodells zu erzielen. Für das Training wurde eine Kombination von Transkriptionen der Sprachaufzeichnungen aus mehreren Korpora verwendet.

Auch Hain und Wan (2006, S. 1069-1072) analysieren die Nutzung von textuellen Daten, die durch den Einsatz von Websuchmaschinen für den Aufbau eines domänenspezifischen Sprachmodells zur Steigerung der Leistungsfähigkeit des

[2] http://www.mit.edu/

Systems gesammelt werden. Hervorzuheben ist die Bedeutung der richtigen Formulierung von Suchanfragen. Davon hängt es ab, welche Daten gesammelt werden, was die Leistungsfähigkeit des Modells, bezogen auf Perplexität und WER[3], entscheidend beeinflusst.

In der Literatur finden sich Hinweise auf die Vorteile, die der Aufbau von mehreren themenspezifischen Sprachmodellen mit sich bringt, die nach dem Training mit dem Basismodell gemixt werden (vgl. Jurafsky & Martin, 2009, S. 147; Gildea & Hofmann, 1999, 2167-2170). Das wird damit begründet, dass Dokumente, die ähnliche Themen behandeln, auch ähnliche Wörter benutzen. Als Konsequenz ergibt sich, dass separate Sprachmodelle für unterschiedliche Themen trainiert werden müssen. Zum Aufbau von themenspezifischen Sprachmodellen empfehlen Sethy, Georgiou und Narayanan (2005, S. 1293–1296) ein iteratives Vorgehen zum Durchsuchen des Web und dem Sammeln von relevanten themenbezogenen Textdaten. Hierfür wurde ein Klassifikationssystem auf Basis von TF/IDF entwickelt, um die Ergebnisse der Suchanfragen zu gewichten und für ein Thema hochrelevante Dokumente herauszufiltern.

Konchady (2008, S. 80) beschreibt TF/IDF als ein Messverfahren[4], mit dem Dokumente klassifiziert und gewichtet werden können. Die erste Komponente TF (*Term Frequency*) gibt die Anzahl des Vorkommens des Begriffs innerhalb eines Dokuments an. Die zweite Komponente IDF (*Inverse Document Frequency*) stellt die Umkehrung der Anzahl der Dokumente in der Gesamtsammlung dar, in denen das Wort mindestens einmal vorkommt. Die Kombination dieser beiden Komponenten wird zur Gewichtung eines Begriffs verwendet.

Vor dem Hintergrund der geschilderten Ansätze zur Domänenadaption eines Spracherkennungssystems wird im nächsten Abschnitt der Bezugsrahmen dargestellt und die Leitfrage und die Zielsetzung für diese Arbeit formuliert.

[3] Die Evaluationsmaße der Spracherkennung werden im Kapitel 4 beschrieben.
[4] Dieses Verfahren wird im Kapitel 5.1.1 dargestellt

1.2 Zielsetzung der Arbeit

An der Technischen Informationsbibliothek[5] (TIB) in Hannover wird seit 2011 im Kompetenzzentrum für nicht-textuelle Materialien[6] (KNM) ein innovatives, medienspezifisches, audiovisuelles Portal entwickelt. Das Ziel ist, wissenschaftliche hochqualitative multimediale Objekte passend zum Fächerspektrum der TIB systematisch zu beschaffen, zu sammeln und langfristig zu archivieren, mittels Metadaten zu beschreiben und durch die Anbindung am GetInfo-Portal[7] für Wissenschaft und Lehre auffindbar und nutzbar zu machen. Aus den naturwissenschaftlichen und technischen Fachgebieten werden Vorlesungsvideos, Videodokumentationen, Visualisierungen von Sachverhalten aus Forschungsberichten, Forschungsdaten, 3D-Objekte und Simulationen gesammelt. Darüber hinaus wurde die Sammlung des Instituts für Wissenschaftlichen Film (IWF)[8] in Göttingen übernommen.

Mit Techniken der Audioanalyse können komplette Transkriptionen automatisch erzeugt und dazu genutzt werden, ein text-basiertes Retrieval der gesprochenen Sprache zu ermöglichen. Für die Audioanalyse wird die automatische Spracherkennung der Firma European Media Laboratory[9] eingesetzt. Der Prozessablauf der Audioanalyse wird in der Abbildung 1 dargestellt. Das installierte Vokabular hat einen Umfang von ca. eine Million Wortformen. Die Spracherkennung unterstützt die Sprachen Deutsch und Englisch. Anfang 2012 wurde die automatische Spracherkennung evaluiert und die Erkennungsgenauigkeit ermittelt. Bei zwei durchgeführten Untersuchungen lagen die ermittelten Durchschnittswerte für WER zwischen 49,4% und 53,2%.[10]

[5] http://www.tib-hannover.de
[6] http://www.tib-hannover.de/de/dienstleistungen/kompetenzzentrum-fuer-nicht-textuelle-materialien-knm/
[7] http://www.tib-hannover.de/de/getinfo/
[8] http://de.wikipedia.org/wiki/IWF_Wissen_und_Medien
[9] http://www.eml-development.de/
[10] Die Angaben sind einem mir vorliegenden internen Dokument entnommen.

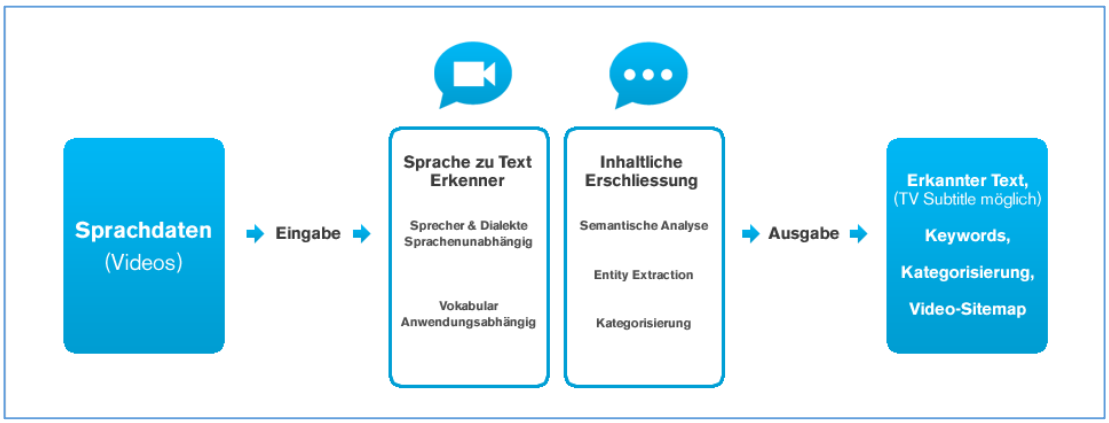


Abb. 1-1: Prozessablauf bei der Audioanalyse. Quelle: http://vetail-x.com/transkription

Die leitende Fragestellung dieser Arbeit ist: Kann der prozentuale Anteil der korrekt erkannten Wörter durch Training des Sprachmodells der automatischen Spracherkennung mit fachspezifischer Terminologie wirksam gesteigert werden?

Mit der Fragestellung sind folgende Ziele verknüpft: Darzulegen, welche Anzahl und welche Art von Daten benötigt werden, um den Prozentsatz der falsch erkannten Wörter zu senken. Zu untersuchen, wie viel fachspezifisches Vokabular in Vorlesungsvideos im Vergleich zu wissenschaftlichen Publikationen benutzt wird um Unterschiede hinsichtlich der verwendeten Fachsprache aufzuzeigen. In diesem Zusammenhang soll gezeigt werden, wie Fachterminologie erkannt werden kann und welche Besonderheiten es bei technischen Fachbegriffen gibt.

1.3 Methodische Herangehensweise

Für die Bearbeitung der Fragestellung stellte das Kompetenzzentrum für nicht-textuelle Materialien der Verfasserin die Möglichkeit zur Verfügung, mit der web-basierten Komponente der Transkriptionsplattform *Language Model Workplace*[11] des European Media Laboratory zu arbeiten. Die integrierten Tools ermöglichen es, das Sprachmodell mittels spezifischer Daten aus dem Web sowie Textkorpora anzupassen, das Domänen-Training nach Projekten zu differenzieren, Trainings- und Testdaten zu definieren, eine Adaption des Sprachmodells und Tests durchzuführen sowie die Ergebnisse anzuzeigen.

[11] Die URL von Language Model Workplace darf nicht veröffentlicht werden.

Zum Training des Sprachmodells wird in Anlehnung an Fügen (2009, S. 67–90) ein kombiniertes Verfahren eingesetzt. Zuerst werden über das GetInfo-Portal Vorlesungsvideos in der Sammlung von KNM nachgewiesen und die Namen von Rednern ermittelt. Aus dem Titel der Vorlesung oder aus den vorhandenen Präsentationsfolien wird das Thema ermittelt. Dies bildet die Grundlage für die Ableitung von Suchbegriffen, mit denen nach relevanten Webseiten und Transkriptionen, fachspezifischen Publikationen und wissenschaftlichen Videos gesucht wird. Alle relevanten Quellen werden für das Training des Sprachmodells in Language Model Workplace verwendet.

Für die Beantwortung der Untersuchungsfrage zu den Unterschieden zwischen den Fachsprachen in Vorlesungsvideos und wissenschaftlichen Publikationen wird eine Sammlung von fachspezifischen Texten aufgebaut, aus denen die Fachterminologie extrahiert wird. Zur Identifizierung und Extraktion von Fachbegriffen wird die Differenzanalyse eingesetzt, ein Text-Mining-Verfahren. Sie ermöglicht die Identifikation von Schlüsselwörtern und Wortkombinationen, indem die Häufigkeiten des Vorkommens der Wortformen innerhalb von zwei Textkorpora verglichen werden. Für dieses Verfahren wird deshalb neben dem Analysekorpus, der die Fachtexte enthält, auch ein allgemeinsprachiges Textkorpus benötigt (Heyer, Quasthoff & Wittig, 2006, S. 95).

1.4 Aufbau der Arbeit

Im Kapitel 2 wird zunächst der Ausdruck *Begriff* definiert und es werden die Unterschiede zwischen Begriff, Benennung und Konzept sowie die Relation zwischen diesen Ausdrücken und *Terminus* dargestellt. Danach werden die Charakteristika von Fachsprachen und zwei Verfahren beschrieben, die bei der Bildung von Termini verwendet werden. Welches die Besonderheiten der technischen Terminologie sind und nach welchen Mustern sie gebildet wird, wird im letzten Teil des Kapitels gezeigt.

Im Kapitel 3 wird die methodische Vorgehensweise bei der Suche nach Informationen dargestellt: Vorbereitung der Recherche, Aufbau von Recherchestrategien und die Methode für die Formulierung von Suchanfragen. Danach werden die Typen der verwendeten Rechercheinstrumente mit Beispielen beschrieben.

Im Kapitel 4 werden zunächst die unterschiedlichen Techniken zur Adaption von Sprachmodellen dargestellt. Im Anschluss daran werden die Rahmenbedingungen und das verwendete Verfahren bei der in dieser Arbeit durchgeführten Untersuchung zur Adaption eines Sprachmodells beschrieben. Danach werden die für das Training des Sprachmodells und das Testen der Adaption verwendeten Datentypen, deren Umfang und die darin behandelte Thematik beschrieben. Daran anknüpfend werden die durchgeführten Experimente und ihre Ergebnisse beschrieben.

Im Kapitel 5 werden zunächst ausgewählte statistische, linguistische und hybride Methoden zur Terminologie-Extraktion dargestellt, anschließend die in dieser Arbeit verwendete Methode, das Programm zur Terminologie-Extraktion und die vorbereitenden Arbeiten am Textkorpus. Mit der Darstellung der Ergebnisse der Terminologie-Extraktion wird das Kapitel 5 abgeschlossen.

Mit einer Zusammenfassung und einem Ausblick im Kapitel 6 wird die Arbeit abgeschlossen.

2. Terminologie und Fachsprache

Die Terminologie wird von Felber und Budin (1989, S. 1–5) als eine geordnete Menge von Begriffen mit den dazugehörigen Begriffszeichen eines Fachgebietes definiert. Begriffe stellen Abbildungen konkreter oder abstrakter Gegenstände in unserer geistigen Vorstellung dar. Die Begriffszeichen sind Schriftzeichenfolgen, die den Begriffen zugeordnet werden und zur Benennung, Verschriftlichung und Versprachlichung der Begriffe dienen. Eugen Wüster, der als einer der Pioniere unter den Terminologieforschern gilt, hat den „Begriff" definiert als „das Gemeinsame, das Menschen an einer Mehrheit von Gegenständen feststellen und als Mittel des gedanklichen Ordnens (‚Begreifens') und darum auch zur Verständigung verwenden" (1979, S. 7 zit. in Felber & Budin, 1989, S. 24). Auch Rothkegel (2010, S. 246) definiert den „Begriff" als eine „abstrakte Vorstellung von einer Klasse von Gegenständen oder Sachverhalten", für die als Synonyme „Konzepte" oder „Denkinhalte" verwendet werden können.

Die Gegenstände haben Eigenschaften, die auch mit Begriffen ausgedrückt werden können. Diese Merkmal-Begriffe stellen als Gesamtheit den Begriffsinhalt dar. Die Beschreibung eines Begriffs erfolgt durch die Angabe aller Merkmale des Begriffsinhaltes und des Begriffsumfanges. Dadurch wird der Begriff definiert und somit auch der Gegenstand beschrieben (vgl. Felber & Budin, 1989, S. 1-5).

In der Terminologie werden die Begriffe als unabhängig von deren Benennungen (Termini) angesehen. Die Benennung eines Begriffs wird nach Felber und Budin (1989, S. 41) als „fachliches Lexem betrachtet, die Terminologie als Subsprache". Das Lexem wird dort definiert als eine Konstruktion von unterschiedlichen bedeutungstragenden Elementen der Sprache, die eine Einheit darstellen. Die Begriffe eines Sachgebietes stehen in unterschiedlichen Relationen zueinander und bilden gemeinsam ein Begriffssystem (vgl. Felber & Budin, 1989, S. 32–36). Die Terminologien spiegeln als Begriffssysteme das Wissen des jeweiligen Sachgebietes wider und gehören als Teilsprachen der Allgemeinsprache an (vgl. Felber & Budin, 1989, S. 41).

In der Deutschen Industrie-Norm 2342 (vgl. Schmalenbach, 2000) wird der Begriff „Fachsprache" definiert als ein „Bereich der Sprache, der auf eindeutige und

widerspruchsfreie Kommunikation in einem Fachgebiet gerichtet ist und dessen Funktionieren durch eine festgelegte Terminologie entscheidend unterstützt wird.“ Als Pendant zum deutschen Wort „Fachsprache“ sind in der englischen Sprache die Ausdrücke „English for special purposes (ESP)“ und „Special language“ üblich (vgl. Sager, Dungworth & McDonald, 1980, S. 1–5). Im Sinne der DIN-Definition besteht die Aufgabe von Fachsprachen darin, einen Bestand von Termini bereitzustellen, die der internationalen Fachcommunity ermöglichen, „präzise und ökonomisch“ über einen Sachverhalt zu kommunizieren. Die Präzision wird durch die Eindeutigkeit der Termini gewährleistet. Die Eindeutigkeit der Begriffe ermöglicht einen Austausch über fachliche Konzepte unabhängig von der Sprache (vgl. Witschel, 2004, S. 14).

Nach Witschel (2004, S. 15) unterscheiden sich Fachsprachen von der Allgemeinsprache durch die signifikante Menge von Fachtermini. Auf der lexikalischen Ebene sehen auch Ahmad und Rogers (1997, S. 731) den entscheidenden Unterschied zwischen Fach- und Allgemeinsprache: Das Domänenwissen ist in Form von Termini in den Fachpublikationen der Experten dokumentiert. Das domänenspezifische Vokabular wird auf Basis der Wortformen aus der Allgemeinsprache, aus Terminologien anderer Wissensgebiete und aus Wortformen gebildet, die weder in der Allgemeinsprache noch in anderen Fachgebieten üblich sind. Ahmad und Rogers (1997, S. 731) bezeichnen die letzteren als „the most clearly identifiable as terms.“

Neben Witschel (2004, S. 15) weisen auch Ahmad und Rogers (1997, S. 725–731) auf die spezielle Syntax hin, die von Wissenschaftlern in den fachspezifischen Publikationen bevorzugt verwendet wird. Charakteristisch für die Fachsprachen ist die hohe Anzahl von Nominalisierungen, Komposita, Phrasen und Pluralformen sowie Passivkonstruktionen, die insbesondere in technischen Fachsprachen vorkommen. Bezugnehmend auf „English Special Languages“ (vgl. Sager et al., 1980) weisen Heyer, Quasthoff und Wittig (2006, S. 46) auf die fachgebietsspezifische Morphologie hin, welche, durch die signifikant häufige Nutzung charakteristischer Morpheme und Derivationen, die wissenschaftlich-technischen Fachsprachen kennzeichnet. Das Morphem ist das kleinste bedeutungstragende Element der Sprache. In den Fachsprachen finden sich Morpheme

in Form von Suffixen und Präfixen, z.B. *-itis* in medizinischen Fachbegriffen, wie bei Mening*itis*. Als Derivationen oder Derivative bezeichnet man neu durch die Anfügung von Morphemen an die Wortstämme erschaffene Wortformen, z.B. *lösbar* durch Anfügung des Morphems *bar* (vgl. Heyer et al., 2006, S. 47; S. 327).

Die Kernfrage lautet, wann ist ein Wort ein Fachterminus und wie kann man Fachtermini in Texten erkennen. Nach Witschel (2004, S. 16) gibt es zwei wichtige Eigenschaften von Termini, die ihre Identifizierung ermöglichen: Sie werden nach bestimmten Mustern gebildet und, verglichen mit Wörtern der Allgemeinsprache, signifikant häufig im Text verwendet. Die linguistischen Merkmale der Fachtermini, die ihre Erkennung im Text ermöglichen, werden im nächsten Abschnitt dargestellt.

2.1 Bildung von Fachtermini und deren Charakteristika

Fachtermini sind die fachlichen Ausdrücke, die in bestimmten Wissenschaftsgebieten bewusst verwendet werden, um eine effektive Kommunikation über die Konzepte des jeweiligen Kontexts zu gewährleisten. Dubuc und Lauriston (1997, S. 80–81) nennen zwei Bedingungen für die Beurteilung eines Terminus. Zum einen muss ein Terminus einen Bezug zu einem speziellen Bereich des menschlichen Handelns haben, um als Fachbegriff zu gelten, zum anderen muss der referenzierte Aktivitätsbereich eine spezialisierte Sprache verwenden. Die Zugehörigkeit eines Begriffs zu einem speziellen Bereich kann anhand folgender Merkmale bestimmt werden:

- Der Begriff hat eine linguistische Form, die nur in diesem Fachgebiet vorkommt (*field-specific terms*)
- Der Begriff existiert zwar auch in der Allgemeinsprache sowie in anderen Fachsprachen, im Referenzbereich hat der Begriff allerdings eine spezielle Bedeutung (*specialized meanings*)
- Der Begriff gehört zu einem anderen Fachgebiet, das mit dem jeweiligen Referenzbereich in einer Relation steht (*associated terms*).

Obwohl das Verfahren der Extraktion von Fachtermini aus Textkorpora automatisch oder semi-automatisch abläuft, sind dennoch genaue Kenntnisse über die Muster, die der Begriffsbildung zugrunde liegen, notwendig, um richtige Termini aus der Menge der extrahierten potentiellen Begriffe zu identifizieren. Die etablierten Methoden zur Begriffsbildung werden nachfolgend geschildert. Nach Sager (1997, S. 25) stellt die Begriffsbildung einen intellektuellen Prozess dar, während dessen die Benennung von Konzepten eines speziellen Fachgebietes unter Berücksichtigung existierender Muster erfolgt. Sager nennt drei der verwendeten Methoden bei der Bildung von Fachtermini:

- Die existierenden Termini werden verwendet, um ein neues Konzept zu definieren
- Die existierenden Sprachressourcen werden verändert, um neue Konzepte zu beschreiben
- Neue linguistische Entitäten (Neologismen) werden erfunden, um neue Konzepte zu benennen.

2.1.1 Bildung von Fachtermini durch Nutzung vorhandener Wortformen

Im Fall der Nutzung von existierenden Ressourcen werden Begriffe aus der Allgemeinsprache verwendet und deren Bedeutung durch die Bildung von Analogien ausgedehnt und auf das neue Konzept übertragen (vgl. Sager, 1997, S. 28–29). Beispiele hierfür sind: *a rock-like substance, an L-shaped room*. Bei der Nutzung von Gleichnissen für die Begriffsbildung werden oft Ausdrücke wie *-style, -like, -type* verwendet, z.B. *tooth-like projection*. Auch die Nutzung von Metaphern oder - wie Sager (1997, S. 29) sie nennt - von metaphorischen Kombinationen ist üblich, wenn zwischen den Begriffen Ähnlichkeiten bezüglich Form, Lage und Funktion gegeben sind (z.B. *umbrella cupola*). Dabei kommt es manchmal vor, dass zwischen dem allgemeinen Begriff und dem Fachbegriff nicht unterschieden werden kann. Ein gutes Beispiel dafür ist *anchor* in der Fachsprache der Webprogrammierung.

Eine andere Methode besteht im interdisziplinären Ausleihen[12]: Die Benennung eines Konzeptes in einem Fachgebiet wird für das Konzept eines anderen Bereichs „ausgeliehen“ und erhält dabei eine andere Bedeutung. Das Resultat dieser Methode ist das Erfinden von Homonymen (vgl. Sager, 1997, S. 28–29). Witschel (2004, S. 22) schildert eine weitere Methode, die er „Lehnübersetzung“ nennt, bei der ein fremdsprachiger Terminus wortwörtlich übersetzt wird (z.B. wurde *Luftbild* aus der wörtlichen Übersetzung von *air photo* gebildet).

2.1.2 Bildung von Fachtermini durch Veränderung vorhandener Wortformen

Um neue Begriffe aus existierenden Wortformen zu bilden, indem sie verändert werden, können verschiedene Methoden verwendet werden. In der Fachliteratur werden folgende genannt: Derivation, Komposition, Phrasenbildung, Konversion und Kompression (vgl. Sager, 1997, S. 30–38; Heyer et al., 2006, S. 269–272; S. 327–330).

Derivation

Bei der Derivation wird einem Wortstamm eine andere Wortform als Suffix oder Präfix angefügt. In den englischen Fachsprachen ist die Menge der Affixe[13] größer als in der Allgemeinsprache, da zahlreiche Wortstämme und Affixe aus Latein und Griechisch übernommen und angepasst wurden. Ein Beispiel: *telematics* wurde aus dem griechischen Wortelement *tele* als Präfix und *matics* gebildet. Die Übernahme, Anpassung und Integration von Wortelementen und von Methoden der Wortbildung aus den altklassischen Sprachen ist für die englischen Wissenschaftssprachen kennzeichnend. Insbesondere die naturwissenschaftlichen und technischen Fachsprachen verwenden kontinuierlich und fast ausschließlich lateinische und griechische Wortelemente, um neue Wortformen mittels Affigierung[14] zu bilden. Insbesondere die griechische Sprache, die über ein regelmäßiges, aber kraftvolles System von Suffixen verfügt, wurde von den englischen Fachsprachen umfassend genutzt (vgl. Sager et al., 1980, S. 246, 258-264).

---

[12] Übersetzung der Verfasserin. Witschel (2004, S. 21) verwendet das Wort „Entlehnungen“.

[13] Präfixe und Suffixe sind die sogenannten Affixe (vgl. Sager, 1997, S. 31).

[14] Affigierung bedeutet das Anfügen von Affixen an Wortstämme (vgl. Heyer et al., 2006, S. 327)

Grundsätzlich werden die Präfixe im Englischen zur näheren Bestimmung von Konzepten (z.B. *super*elevation), zur Bildung von Konzeptpaaren (z.B. *up*stream - *mid*stream - *down*stream) und zur Bildung von Gegensätzen verwendet, wie z.B. bei *un*load, *dis*connect, *de*compose, *in*efficient, *a*nhydrous (vgl. Sager, 1997, S. 31–34).

Durch die Anfügung von Suffixen wird die Wortklasse des Wortelements geändert. Die Suffixe *-age, -ate, -esce, -ize, -ise* und besonders häufig *-ify* werden für die Bildung von Verben aus Substantiven oder Adjektiven eingesetzt, um einen Prozess auszudrücken (z.B. acid*ify*, standard*ize*). Verben werden in Substantive mit den Suffixen *-ing, -ion, -ation* geändert (z.B. plan - plann*ing*, irrigate - irrig*ation)*. Auch Adjektive werden mittels Suffixen in Substantive geändert. Hierfür werden folgende Wortformen verwendet: *-age, -ance, -cy, -escence, -ity*. In den Texten aus dem Fachgebiet Ingenieurwissenschaften wird der Suffix „-ity" zur Beschreibung einer Eigenschaft verwendet (z.B. intense – intens*ity*). Es gibt außerdem eine Reihe von Suffixen zur Formung von Adjektiven: *-al, -ant, -ar, -ary, -en, -ent, -ic, -ile/ible, -ive, -lent, -ly, -ory, -ous, -y* (z.B. angul*ar*, explorat*ory*, flex*ible*).

In der deutschen Sprache wird häufig *-er* als Suffix verwendet, speziell in den technischen Fachsprachen, z.B. *Schweißer, Verstärker, Absorber*. Häufige Verwendung als Suffixe in den naturwissenschaftlich-technischen Fachsprachen finden weitere Morpheme wie z.B. *-grad, -heit, -tum, -ial, -ik, -tät, -ator* und *-ase*. Und als Präfixe sind häufig die Morpheme *anti-, mega-, mikro-, multi-, radial-* und *semi-* anzutreffen (vgl. Witschel, 2004, S. 22; Heyer et al., 2006, S. 47, S. 241).

Komposition und Phrasenbildung

Bei der Komposition handelt es sich um die Bildung von neuen Worteinheiten aus mehreren vorhandenen Wortformen, indem sie aneinander gefügt werden. Die neu erschaffenen Worteinheiten erhalten eine neue Bedeutung, unabhängig von der Bedeutung der konstituierenden Wortelemente. In einem Kompositum aus zwei Elementen spezifiziert das erste Element das zweite, das Kernelement[15] genannt wird. Das Kernelement zeigt die Kategorie, zu der das Konzept gehört, und das bestimmende erste Element zeigt das Kriterium zur Bestimmung der

[15] Übersetzung der Verfasserin. Witschel (2004, S. 22) verwendet hierfür die Bezeichnung „Kopf".

Unterkategorie, z.B. *wind load, ice load, marble structure, metal structure* (vgl. Sager, 1997, S. 31–37).

Komposita selbst können mit anderen Begriffen kombiniert werden, um neue Termini zu bilden (z.B. *rock-type soil*). Dabei können sehr komplexe Komposita entstehen, deren Bedeutung nur im Zusammenhang mit dem speziellen Themenbereich offensichtlich wird (z.B. *data control block fill-in process, minimum strain energy twisted-folded energy*). An den Beispielen wird sichtbar, dass im Englischen die Beziehung zwischen den Elementen des Kompositums mithilfe eines Bindestrichs verdeutlicht wird (vgl. Sager, 1997, S. 31)**.** Abhängig von der Art des Kernelementes wird unterschieden zwischen folgenden Kompositatypen:

- Komposita zur Beschreibung von Objekten, z.B. *concrete roof*
- Komposita zur Beschreibung von Eigenschaften, z.B. *concrete stability*
- Komposita zur Beschreibung von Prozessen oder Aktivitäten, z.B. *concrete slump test, concrete casting* (vgl. Sager, 1997, S. 35)**.**

Kompositabildung führt im Englischen zur Bildung von Mehrworttermini.

Anders verhält es sich in der deutschen Sprache. Zur Kompositabildung werden zwei Wortstämme aneinandergefügt (vgl. Witschel, 2004, S. 22)**.** Daraus entsteht ein Einwortbegriff. Ähnlich wie in der englischen Sprache gibt es auch bei deutschen Komposita ein Kernelement, das die Art des Begriffes bestimmt und ein modifizierendes Element, das die Bedeutung des Begriffs einschränkt. Bei der Kompositabildung werden häufig folgende Muster beobachtet:

- Zwei Nomina wurden zusammengesetzt, z.B. *Konjunkturbarometer*
- Ein Nomen wurde mit einem Verb zusammengesetzt, z.B. *farbabweisend*
- Ein Adjektiv wurde mit unterschiedlichen Kernelementen zusammengesetzt, z.B. *Leichtöl, feingemahlen*.

Nach Sager (1997, S. 36) ist die Phrasenbildung sprachabhängig und ähnelt der Kompositabildung. Er fügt hinzu, dass die germanischen Sprachen eher dazu prädestiniert sind, Komposita zu bilden - verglichen mit den romanischen Sprachen. Es liegt die Vermutung nahe, dass sich Sager auf das Zusammenfügen von mehreren Wortformen zu einer neuen Wortformeinheit bezieht, was in den germanischen Sprachen zur Kompositabildung üblich ist. Diese Vermutung der

Verfasserin wird von Heid (1997, S. 790–791) bestätigt, der den Terminus „noun compounds“ mit Deutsch, Finnisch und Holländisch in Verbindung bringt und die gleiche These wie Sager (1997, S. 36) postuliert. Heid erweitert jedoch die These und beschreibt die Auswahl von Kompositaelementen, welche aus Nomen gebildet werden, als „ [...] clear combinatory preferences, often merely conventional, that in many cases go as far as the complete terminological ‚fixing' oft the compounds and noun groups“ (vgl. Heid, 1997, S. 791).

Auch Wright (1997, S. 15) weist auf die Phrasenbildung aus mehreren Nomen hin, die sie als „free-formed combinations“ bezeichnet (z.B. *gold and foreign currency reserves*). Der Grund dafür ist, dass solche Konstrukte weder zu Phrasen noch zu Komposita gehören, aber häufig zusammen erscheinen. In solchen Fällen muss geprüft werden, ob die Benennung des Konzepts in anderen Sprachen als Mehrwortterminus vorkommt.

Als Vorteil von Phrasenbildung zur Benennung von Konzepten sieht Sager (1997, S. 37) die differenzierte Darstellung von unterschiedlichen Merkmalen (z.B. *step-by-step variable speed transmition*). Werden bei der Phrasenbildung Präpositionen, Konjunktionen, Artikel und Adverbe verwendet, wird ein Bindestrich eingesetzt, um die Beziehungen zwischen den Elementen zu verdeutlichen, z.B. *stick-and-rag work*.

Sager (1997, S. 36) erklärt weiterhin, dass im Englischen ein Phrasenkonstrukt gebildet wird, falls ein Kompositum das Konzept nicht eindeutig genug bezeichnen kann. Er veranschaulicht anhand vieler Beispiele, wie ein Kompositum, meistens aus zwei Wortformen, in der englischen Sprache gebildet wird. Dabei erwähnt er auch die Möglichkeit, ein Kompositum mit weiteren Wortformen zu kombinieren und zeigt verschiedene Muster dafür. Obwohl Sager (1997, S. 34) die aus mehreren Komposita neugeschaffenen Konstrukte immer noch Komposita nennt, wird der Unterschied zu den Phrasentermini nicht sichtbar. Das erste Beispiel zeigt die Zusammensetzung zweier Komposita, das zweite stellt einen Phrasenterminus dar (vgl. Sager, 1997, S. 34, S. 37):

- reinforced concrete underwater construction
- infinitely variable speed transmission.

Dass die Grenze zwischen solchen Komposita und Mehrworttermini schwierig zu beschreiben ist, hebt erneut Heid (1997, S. 791) hervor und führt diesen Zustand auch auf die fehlenden Standards zu dieser Thematik zurück.

Von Bedeutung erscheinen in diesem Zusammenhang die Kollokationen (auch Kookkurrenzen genannt), die als Kombinationen von Phrasen-Termini definiert werden (vgl. Wright, 1997, S. 15–16; Witschel, 2004, S. 24–25**)**. Kennzeichnend für die Kollokationen ist, dass sie signifikant häufig in einem Satz, Absatz und sogar in unterschiedlichen Absätzen zusammen vorkommen und dabei das gleiche Muster immer wieder verwendet wird. Wright (1997, S. 15) präsentiert dort die Meinung zweier Expertengruppen: einerseits werden unterschiedliche Typen von Mehrwort-Strings als Kollokationen bezeichnet, andererseits sollen nur „combinatory phraseological units" Kollokationen genannt werden, welche „are made up of words that frequently co-occur (i.e., „co-locate"), even if they don't always follow each other directly in sentences". Eine präzisere Definition formuliert Heid (1997, S. 788–789), die besagt, dass Kollokationen *zwei* Lexeme beinhalten, welche in einer spezifischen Relation zueinander stehen, nach bestimmten Mustern gebildet werden und eines der Lexeme ein Konzept bezeichnet. Auch Manning und Schütze (2001, S. 183) weisen auf die Verwendung von unterschiedlichen Definitionen[16] für das Konzept von Kollokationen hin und bemerken zurecht, dass „the notion of collocation may be confusing to readers without a background in linguistics".

Es gibt drei wichtige Kriterien für die Identifizierung und linguistische Bearbeitung von Kollokationen (vgl. Manning & Schütze, 2001, S. 184):

- Sie haben eine feste Bedeutung, die unabhängig von der Bedeutung der konstituierenden Teile ist.
- Die Bestandteile von Kollokationen können nicht durch andere mit ähnlicher Bedeutung ersetzt werden.
- Sie können durch Hinzufügung zusätzlicher Begriffe nicht verändert werden.

[16] Im Kapitel 5.1.1 wird gezeigt, wie die Begriffe Kookkurrenzen und Kollokationen von unterschiedlichen Autoren gleichermaßen definiert werden.

Mit fachspezifischen Kollokationen können komplexe Sachverhalte dargestellt und Theorien formuliert werden. Die fachspezifischen Kollokationen werden aus mehreren Konzepten gebildet, die in einer wechselseitigen Relation stehen.

Konversion

Bei der Konversion handelt es sich um die Umwandlung der Wortklasse einer gegebenen Wortform ohne das Anhängen von Derivativen. Üblicherweise werden bei dieser Methode Verben nominalisiert, z.B. *das Schmelzen* (vgl. Witschel, 2004, S. 22; Sager, 1997, S. 37).

Kompression

Bei der Kompression handelt es sich um jede Art von Abschneidung oder Komprimierung eines langen Terminus (vgl. Sager, 1997, S. 37–38). Eine Kompressionsart ist die Bildung von Akronymen, die aus den Initialen der Begriffsbezeichnung entstehen, z.B. UNESCO, NATO. Eine andere Methode ist das Abkürzen eines Terminus, indem einzelne Buchstaben oder mehrere Silben weggelassen werden, z.B. *math, lab*. Sager (1997, S. 37–38) schildert noch ein spezielles Modell, das darin besteht, zwei Methoden miteinander zu kombinieren. Zuerst wird ein Kompositum gebildet, das gleichzeitig im Bildungsprozess Teile (Silben) durch Abkürzen verliert, z.B. *biological* + *electronic* = *bionic*.

2.2 Erkennung der technischen Fachterminologie unter Berücksichtigung ihrer Besonderheiten

Nach Chung und Nation (2004, S. 251–252) besteht die Hauptschwierigkeit in der Entscheidung, ob ein Begriff ein technischer Terminus ist, darin, dass die Technikbezogenheit ein zweckbestimmender Aspekt des Wortes ist und die spezielle Verwendung des Begriffs zusätzlich berücksichtigt werden muss. Das technische Vokabular steht in enger Beziehung mit den behandelten Themen, tritt nur in spezialisierten Fachgebieten auf und ist Teil des jeweiligen Wissenssystems (vgl.

Chung & Nation, 2004, S. 251–252; Justeson & Katz, 1995, S. 10). Diese Merkmale stellen die Basis für die Identifizierung von technischen Fachtermini dar. Um festzustellen, wie hoch die Relation zwischen den Begriffen und dem Themenbereich ist, kann ein Ranking-Verfahren angewendet werden (vgl. Chung & Nation, 2004, S. 252). Die wichtigsten Verfahren zur Identifizierung und Extraktion von Fachtermini werden im Kapitel 5 *Terminologie-Extraktion* dargestellt.

Die Wechselwirkung zwischen der behandelten Thematik in einem Fachtext und dem verwendeten Fachvokabular wurde auch von Justeson und Katz (1995, S. 12) geschildert: Wird ein Konzept mehrmals durch die Verwendung desselben Terminus wiederholt, ist das ein Signal, dass dieses Konzept einen relevanten Aspekt der Thematik darstellt oder zumindest eine wichtige Rolle in der Argumentation spielt. Diese Beobachtung ist bei der Identifizierung von Fachtermini hilfreich. Die Benennungen, die bei der Beschreibung der im Fachtext behandelten Konzepte verwendeten werden, sind somit immer die relevanten Fachtermini für den Themenbereich.

Hilfreich bei der Identifizierung von Fachtermini sind auch die von den Autoren selbst gegebenen Hinweise in den Texten, um die Aufmerksamkeit auf die Botschaften der Abhandlung zu lenken. Solche Hinweise können in Form von Markierungen oder Hervorhebungen angebracht sein, z.B. von wichtigen Begriffen für die präsentierten Thesen und von Neologismen (vgl. Chung & Nation, 2004, S. 252). Oft werden Definitionen oder Synonyme für den hervorgehobenen Begriff angeboten. In grafischen Darstellungen werden die technischen Termini durch eine Aufschrift hervorgehoben. Auch wenn diese typographischen Hinweise nicht immer eindeutig, deshalb schwierig wahrzunehmen sind und auch andere problematische Auswirkungen haben, können sie dennoch als eine hilfreiche Informationsquelle für die Erkennung von Termini in (nicht vorverarbeiteten) fachspezifischen Texten gesehen werden (vgl. Chung & Nation, 2004, S. 256–257).

Bei der Verwendung von technischen Termini in spezialisierten Texten wurden bestimmte Muster beobachtet, die im folgenden Abschnitt dargestellt werden. Nach Justeson und Katz (1995, S. 9–12) bestehen die meisten technischen Begriffe aus komplexen Phrasen, die immer wieder dazu verwendet werden, um

ein Konzept zu referenzieren. In technischen Texten finden sich fast ausschließlich solche terminologischen Einheiten und diese sind meistens lexikalisch, d.h. sie werden in einem Fachwörterbuch nachgewiesen. Gerade die Wiederholung von Mehrworttermini, die durch Synonyme nicht ersetzt werden können, hilft bei der Unterscheidung zwischen Fachterminologie und Allgemeinsprache und trägt zur Identifizierung von Terminologie in Texten bei.

Diese terminologischen Phrasen werden meistens aus Substantiven und Adjektiven gebildet, selten auch mit Präposition (z.B. mit *of*). Der Grund dafür ist, dass den technischen Fachgebieten eine Organisation ähnlich wie Taxonomien typisch ist. Nominalphrasen werden bevorzugt gebildet, um komplexe Entitäten, wie z.B. technische Mittel, Objekte, Sachverhalte, Vorgänge und Verfahren, Ergebnisse sowie charakteristische Eigenschaften und Umstände, mit der notwendigen Eindeutigkeit darzustellen, was mit Einworttermini nicht möglich wäre. Werden Einworttermini verwendet, dann sind diese Begriffe meistens lateinischen oder griechischen Ursprungs. In den technischen Texten ist auch die geringe Verwendung von Adjektiven, Verben und Adverben sowie von Präpositionen und Konjunktionen auffallend (vgl. Justeson & Katz, 1995, S. 13).

Zusammenfassung

Ein Begriff stellt eine bedeutungsvolle sprachliche Einheit dar, die aus einem oder mehreren Wörtern gebildet wird und ein spezifisches Konzept in einem Fachgebiet eindeutig repräsentiert. Die Terminologie umfasst alle einem Fachgebiet zugehörigen Termini und spiegelt die strukturelle Organisation der Konzepte des jeweiligen Fachgebiets wider. Texte, die ausschließlich einem Fachgebiet zugeordnet werden können und spezielle Themen behandeln, verwenden eine spezifische Sprache, die zwar auf der Allgemeinsprache aufbaut, aber sich von ihr in Syntax, Semantik, im verwendeten Wortschatz und in der Wortbildung unterscheidet.

3. Informationsquellen für das Training der automatischen Spracherkennung und den Erwerb von Fachterminologie

Nach Sühl-Strohmenger (2008, S. 33–35) wächst im digitalen Zeitalter die Bedeutung der wissenschaftsrelevanten Information, verglichen mit der klassischen wissenschaftlichen Fachinformation, stetig an. Der Grund dafür wird im schnellen Wissensgenerierungsprozess gesehen. Wissenschaftliche Forschung ist jedoch ohne uneingeschränkten Zugang zu vielfältigen Informationsressourcen, wie sie ehemals nur in wissenschaftlichen Bibliotheken und Informationseinrichtungen gesammelt und bereitgestellt wurden, nicht denkbar.

Wissenschaftsrelevante Informationen und Wissen, noch in der Entstehung oder als „Endprodukt“, werden nicht mehr ausschließlich über die klassischen Verlagsprodukte veröffentlicht, sondern vermehrt über alle zur Verfügung stehenden webbasierten Kommunikationskanäle, wie Mailinglisten, Weblogs, Wikis, Repositories, Forschungskommunikationsplattformen und über die Webseiten von Autoren oder Fakultäten. Das Internet ist zu einer bevorzugten Quelle von wissenschaftlich relevanten Informationen geworden, was auch für die vorliegende Aufgabenstellung genutzt wurde, zusammen mit den digitalen Informationsressourcen der Technischen Informationsbibliothek[17] in Hannover.
Für das Training des Sprachmodells der automatischen Spracherkennung und für den Aufbau einer Dokumentensammlung für die Terminologie-Extraktion wurden Recherchen mit dem Ziel durchgeführt, wissenschaftliche fachspezifische und thematisch relevante Webseiten und digitale Volltexte nachzuweisen. Hierfür wurden unterschiedliche Rechercheinstrumente und Informationsquellen berücksichtigt, jedoch wurde der Schwerpunkt auf webbasierte Datenbanken und auf spezialisierte Internet-Suchdienste gelegt.

[17] http://www.tib.uni-hannover.de/de/literatursuche/fachdatenbanken.html

3.1 Methodische Vorgehensweise

Die Suche nach Informationen bedarf nach Borgman (1996, S. 493–503, zit. in Armstrong & Large, 2001, S. 1–2) vielfältiger Kenntnisse und Fähigkeiten. Zuerst müssen der Informationsbedarf erkannt und die Begriffe abgeleitet werden, die in eine Fragestellung umgewandelt werden. Semantisches Wissen ist nach Borgman notwendig, um aus den Suchbegriffen eine logische Suchabfrage in einem Suchsystem zu entwickeln. Ein gewisses technisches Wissen ist notwendig, um unterschiedliche Suchsysteme zu bedienen und die Suchanfragen entsprechend den Systemanforderungen umzusetzen.

Die Vorbereitung der Recherche beginnt mit der Umformulierung des Informationsbedarfs in eine Recherchestrategie (vgl. Poetzsch, 2006, S. 24). Der Informationsbedarf ergibt sich aus der Zielsetzung der Arbeit, die darin besteht, die Erweiterung des Vokabulars und die Veränderung des Sprachmodells der automatischen Spracherkennung durch die Hinzufügung von fachspezifischen Texten zu erreichen. Zu diesem Zweck wurden aus den Fächern Informatik und Technik zunächst einige Videos mittels einer Suche im GetInfo-Portal ausgewählt und der Name des Sprechers sowie das behandelte Thema ermittelt. Um das Thema zu ermitteln, wurden Hörproben und die im System gespeicherten inhaltlichen Angaben verwendet. Um Zugang zu den ermittelten Themen zu finden, wurden Vorabrecherchen mit einer allgemeinen Suchmaschine durchgeführt.

Der Aufbau einer Recherchestrategie beinhaltet die Auswahl der Methoden für die Recherche und für die Formulierung der Suchanfragen. Abhängig von den Suchmöglichkeiten des verwendeten Rechercheinstruments, wurde die Suche mit kontrolliertem Vokabular bevorzugt eingesetzt. Die Suche mit Deskriptoren ermöglicht eine präzise und effiziente Suche, da auch Dokumente gefunden werden, in denen Synonyme oder Quasisynonyme für die Darstellung der jeweiligen Konzepte verwendet werden. Die Bereitstellung von kontrolliertem Vokabular in Form eines Thesaurus gibt eine Übersicht über die Hierarchien bzw. Relationen, die zwischen den Begriffen bestehen und ermöglicht dadurch eine Erweiterung oder Einschränkung der Suche. Die Suche mit Deskriptoren bedarf des Fachwissens und steht in engem Zusammenhang mit der Kenntnis von Fachtermi-

nologie. Von der Auswahl der Suchbegriffe hängen Menge und Relevanz der Ergebnisse ab (vgl. Armstrong & Large, 2001, S. 2; S. 11–13).

Die zweite Recherchemethode stellt die Freitextsuche dar, die immer eingesetzt werden soll, wenn es kein kontrolliertes Vokabular gibt, es keine angemessenen Deskriptoren gibt oder keine relevanten Treffer bei der Suche mit Deskriptoren erzielt werden. Die Vorteile einer Suche mit nicht normierten Begriffen liegen darin, dass eine größere Anzahl von Wortkombinationen zur Bildung von Suchanfragen möglich ist und die verwendeten Begriffe der natürlichen Sprache in den Dokumenten gefunden werden können (vgl. Armstrong & Large, 2001, S. 11).

Abhängig von den Rechercheergebnissen wurde auch die Methode *Pearl Growing* eingesetzt, um zusätzliche relevante Dokumente zu identifizieren. Bei dieser Methode wird zuerst eine Suche mit einem passenden Suchbegriff durchgeführt, dabei werden einige wenige, aber relevante Treffer ausgewählt und die ihnen zugewiesenen Schlagwörter analysiert. Die Indextermini, die für die eigene Fragestellung als relevant eingeschätzt werden, können anschließend für die Erstellung von Suchanfragen verwendet werden.

Für die Formulierung der Suchanfragen wurde eine klassische Methode verwendet: die Bildung von Suchanfragen auf Grundlage von Begriffsblöcken. Diese Methode besteht darin, aus der Fragestellung Begriffe abzuleiten, welche die genauen Themen der Fragestellung erfassen. Für jeden Begriff werden einige Synonyme und alternative Suchbegriffe (Ober- oder verwandte Begriffe) ermittelt. Durchgeführt wird die Suche, indem der Hauptsuchbegriff mit Hilfe des Operators OR mit den festgelegten alternativen Suchbegriffen verknüpft wird. Die Abbildung 2 veranschaulicht die zuvor dargestellte Methode.

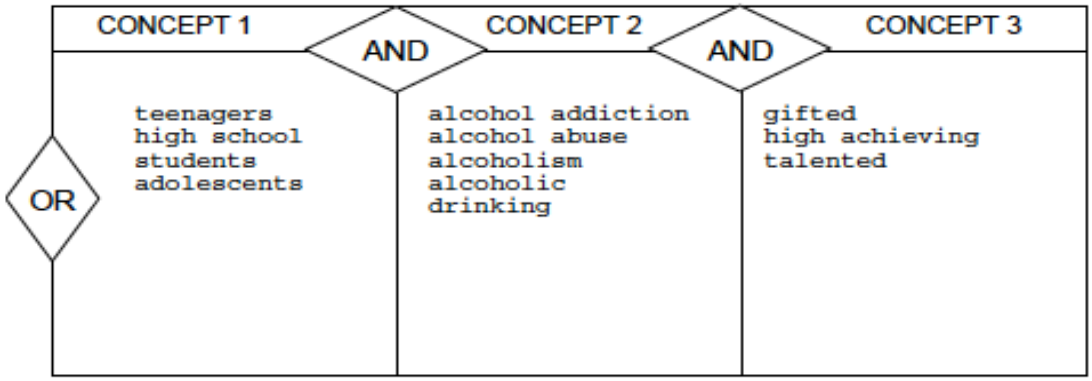


Abb. 2: Begriffsblöcke. Quelle: Dialog Lab Workbook, 2005, S. 2–4

Bei der Formulierung der Suchanfragen wurden die logischen Operatoren AND, OR, NOT verwendet, die unter der Bezeichnung *Boolesche Operatoren* bekannt sind. Trunkierung wurde verwendet, um nach Begriffen mit unterschiedlichen Endungen oder nach Komposita zu suchen. Bei der Trunkierung wird für die Suche ein Wortstamm verwendet, dem direkt dahinter ein Zeichensymbol (meistens wird ein Sternchen- oder Fragezeichen verwendet) angehängt wird. Je nach angebotenen Suchmöglichkeiten des Rechercheinstruments wurden Wildcards verwendet, um nach Begriffen mit unterschiedlichen Schreibweisen zu suchen. Um nach Mehrwortbegriffen zu suchen, wurde die Methode der Phrasenbildung verwendet. Hierfür werden die Suchbegriffe in Anführungszeichen gesetzt, was bedeutet, dass der Mehrwortbegriff als Ganzes und in der angegebenen Reihenfolge gefunden werden muss.

Bei der Auswertung von Rechercheergebnissen wurden die erzielten Treffer nach Typ, Sprache, Aktualität sowie inhaltlicher Relevanz unter Berücksichtigung des Informationsbedarfs analysiert und ausgewählt. Die Relevanz war eines der wichtigsten Kriterien. Da die Verfasserin dieser Arbeit nicht über das notwendige Fachwissen verfügt, um die Relevanz der erzielten Treffer zu beurteilen, wurde diese geschätzt anhand des Übereinstimmungsgrades zwischen den Inhaltsangaben oder den Schlagwörtern, die dem Dokument zugewiesen wurden, und der ursprünglich formulierten Fragestellung. Armstrong und Large (2001, S. 15) bezeichnen diese Art der Relevanzeinschätzung mit „aboutness“. Die genannten Autoren weisen darauf hin, dass die Einschätzung der Relevanz eines gefundenen Dokumentes subjektiv ist und mit dem Informationsbedarf des Suchenden in direkter Relation steht (Armstrong & Large, 2001, S. 15–16).

Es wurden folgende Publikationsarten in englischer Sprache berücksichtigt: Zeitschriften- und Buchaufsätze, Patentschriften und Pre-Prints, technische Berichte und Dissertationen, Konferenzbeiträge und Diskussionsbeiträge in Fachforen und Transkriptionen von gesprochenen Reden. Für die Durchführung der Terminologie-Extraktion wurden sowohl thematisch relevante Volltexte als auch Abstracts ausgewählt, die in digitaler Form vorliegen.

3.2 Typologie der ausgewählten Rechercheinstrumente

Die im Kapitel 3.1 geschilderten Ziele der Arbeit stellten die Basis für die Auswahl der Rechercheinstrumente dar. Um thematisch relevante Publikationen und Webseiten nachzuweisen, wurden geeignete webbasierte Datenbanken, wissenschaftliche Portale und Metasuchmaschinen sowie auf wissenschaftliche Publikationen spezialisierte Suchdienste ausgewählt.

Wichtige Kriterien für die gezielte Auswahl von geeigneten Datenbanken sind Inhalt, Art, Qualität, geographische und zeitliche Abdeckung, Aktualisierungsintervalle, Sprache der Dokumente im Datenbestand sowie Zugriffsmöglichkeiten auf die Datenbanken und die Dokumente selbst (vgl. Poetzsch, 2006, S. 32; Armstrong & Large, 2001, S. 10–11; S. 19–20).
Bei der Auswahl der Suchdienste wurden folgende Kriterien zugrunde gelegt: der fokussierte Gegenstandsbereich und die erfassten Dokumententypen, der Erschließungsgrad von Datenbeständen aus dem *Invisible Web* (dem Teil des World Wide Web, der von Universalsuchmaschinen nicht indexiert wird) und die bereitgestellten Funktionalitäten hinsichtlich Suchanfrageformulierung und Filterung der Ergebnisse (vgl. Griesbaum, Bekavac & Rittberger, 2009, S. 42–46).

Literaturdatenbanken

Die großen Vorteile der Datenbanken liegen in der hohen Qualität und in der präzisen, detaillierten Strukturierung der Informationsinhalte. Neben weit entwickelten und für unterschiedliche Anforderungen anpassbaren Suchmöglichkeiten, wie z.B. die Verwendung von Booleschen Operatoren, von Trunkierung

und Maskierung[18], Nutzung von Feldspezifizierungen und Ähnlichkeitssuche („Find similar“), bieten die Datenbanken Zugriff auf verschiedene Indizes und ermöglichen die Erstellung von komplexen Suchanfragen. Im Ergebnis wird dadurch eine hohe Vollständigkeit (Recall) und Genauigkeit (Precision) der Suchergebnisse (Retrieval) erzielt (vgl. Kind, 2004, S. 389; Sühl-Strohmenger, 2008, S. 190).

Die Informationssuche in fachspezifischen Datenbanken ermöglicht durch die Bereitstellung von Fachthesauri und -klassifikationen und durch die Beschlagwortung mittels fachspezifischer Termini eine gezieltere Suche nach fachlicher Literatur. Die Beschränkung der fachspezifischen Datenbanken auf einschlägige Literatur trägt zu präziseren Rechercheergebnissen bei (vgl. Sühl-Strohmenger, 2008, S. 187–190).

Dass die Fachdatenbanken den Nutzern zahlreiche komplexe Suchmöglichkeiten anbieten, wird von Handreck und Mönich (2008, S. 406) jedoch als einer der Nachteile von Fachdatenbanken gegenüber Wissenschaftssuchmaschinen wie z.B. Google Scholar[19] gesehen, da die Datenbanken in der Bedienung meist nicht intuitiv sind. Ein weiterer Nachteil besteht in der eingeschränkten Zugänglichkeit, da Fachdatenbanken in der Regel lizenzpflichtig sind.

Für die Auswahl der Datenbanken wurden zunächst die Datenbankbeschreibungen im Datenbank-Infosystem[20] (DBIS) verwendet. DBIS wird von der Universitätsbibliothek Regensburg betreut. Es bietet einen zentralen, strukturierten Zugang zum gesamten Datenbankangebot der deutschen wissenschaftlichen Bibliotheken und weist sowohl kostenpflichtige als auch frei verfügbare Datenbanken nach. Die im DBIS nachgewiesenen Informationsressourcen können folgenden Datenbanktypen zugeordnet werden:

- Allgemeine, fachspezifische und fachübergreifende Bibliografien
- Volltextdatenbanken
- Zeitschriftenaufsatzbibliografien

[18] Maskierung bezeichnet die Verwendung von Wildcards, um unterschiedliche Schreibweisen eines Wortes zu suchen (vgl. Kap. 3.1)
[19] http://scholar.google.de/
[20] http://rzblx10.uni-regensburg.de/dbinfo/fachliste.php?lett=l

- Datenbanken, die eine Mischung aus Fachbibliografie und Volltextdatenbank darstellen
- Fachübergreifende Verlagsportale sowie virtuelle Fachbibliotheken.

Alle Datenbanken werden durch formale und inhaltliche Angaben beschrieben und mittels Schlagwörtern und durch Zuordnung zu bestimmten Fachgebieten erschlossen (vgl. Sühl-Strohmenger, 2008, S. 164–168).

Für die Recherche nach relevanten fachspezifischen Publikationen wurden sowohl frei zugängliche als auch lizenzpflichtige Datenbanken ausgewählt. Die Recherche wurde in insgesamt 10 Datenbanken durchgeführt: ENTEC Energietechnik[21], ETDEWEB[22], GetInfo[23], INSPEC[24], SciVerse HUB[25], SpringerLink[26], TEMA[27], ViFaTec[28], Wiley[29], Web of Science[30].
Bei den 10 Datenbanken handelte es sich um sechs bibliographische Datenbanken, zwei fachübergreifende Verlagsportale und zwei fachspezifische, nichtkommerzielle Portale. Vier Datenbanken waren Aufsatzbibliografien. Bei fünf der 10 abgefragten Datenbanken werden nur kurze Zusammenfassungen angeboten. Die anderen fünf Datenbanken bieten Zugriff auf Volltexte bei vorhandener Lizenz. Weiterhin sind fünf von 10 abgefragten Datenbanken lizenzpflichtig. Im Kapitel 3.3 werden die Datenbanken kurz beschrieben und die Auswahlkriterien dargestellt. Eine Übersicht der verwendeten Rechercheinstrumente zeigt die Anlage B1.

Virtuelle Fachbibliotheken

Aus den Bestrebungen der Deutschen Forschungsgemeinschaft zur Verbesserung der Literaturversorgung der Fachwissenschaftler entstand die *Virtuelle Fachbibliothek* als Hybridbibliothek mit dem Ziel, fachrelevante Internetquellen und gedruckte Materialien zu erschließen und zugänglich zu machen. Die Hauptaufgabe der

[21] http://www.wti-frankfurt.de/images/stories/download/de-etec.pdf
[22] https://www.etde.org/etdeweb
[23] https://getinfo.de/app/
[24] http://www.theiet.org/resources/inspec/
[25] http://www.hub.sciverse.com/action/home
[26] http://link.springer.com/
[27] http://www.wti-frankfurt.de/images/stories/download/de-tema.pdf
[28] http://www.vifatec.de/
[29] http://onlinelibrary.wiley.com/
[30] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/

virtuellen Fachbibliotheken besteht in der Bereitstellung eines einzigen Zugangs zu vielfältigen, wissenschaftsrelevanten und fachspezifischen Informationsquellen, unabhängig von Medium, Speicherform und Speicherort. Die virtuellen Fachbibliotheken ermöglichen effizientes Recherchieren durch Verknüpfung vieler Module unter einer Oberfläche.

So fungieren diese Hybridbibliotheken primär als Fachinformationsführer mit einer integrierten Spezialsuchmaschine, die fachlich relevante Internetressourcen erschließt und vollständig indexiert. Sie ermöglichen Fachrecherchen nach gedruckten und elektronischen Literaturquellen durch die Einbindung von relevanten Bibliothekskatalogen der zugehörigen Sondersammelgebietsbibliotheken sowie von deutschen und ausländischen Spezialbibliotheken. Durch die Einbindung von Dokumentlieferdiensten bieten sie die Möglichkeit einer schnellen Literaturbeschaffung für gedruckte Quellen.

Dieses Basisangebot wird auf unterschiedliche Weise durch die Einbindung von bibliographischen und Volltextdatenbanken, elektronischen Zeitschriften, Konferenz-, Fakten- und Adressdatenbanken sowie Tagungskalendern erweitert. Zusätzliche Erweiterungsmodule sind in Form von Tutorials verfügbar, in denen multimediale Lehr- und Lernmaterialien für das Selbststudium bereitgestellt werden, und Newsletter für die regelmäßige Verteilung von aktuellen Informationen (vgl. Rösch & Weisbrod, 2004, S. 177–188; Sühl-Strohmenger, 2008, S. 192–195).

Websuchmaschinen

In Folge der seit Mitte der 90er Jahre anhaltend schnellen Entwicklung von Internet- und Informationstechnologien sind viele Möglichkeiten entstanden, Daten und Informationen im Internet auf einfache Weise zu veröffentlichen.
Mit dem ersten Treffen der *Open Archives Initiative* (OAI)[31] im Oktober 1999 wurde die Open-Access-Bewegung mit dem Ziel gegründet, das Internet als neuen Verbreitungsweg für wissenschaftliche Information verstärkt zu nutzen und

[31] www.openarchives.org

mittels Open Access die nachhaltige Veränderung des wissenschaftlichen Publikationsmarktes zu erzielen. Es galt, die für die publizierenden Wissenschaftler nachteiligen Strukturen und Abläufe des Publikationsmarktes zu vermeiden und damit einen schnellen und unkomplizierten Zugriff auf Forschungsergebnisse und Erkenntnisse anderer Experten, der in vielen Disziplinen unerlässlich ist, zu ermöglichen. Es wurden in Folge viele Dokumentenserver mit Hochschulpublikationen und Open-Access-Repositories aufgebaut und für die Veröffentlichung von Publikationen nach den Prinzipien der Open-Access-Initiative unterschiedliche Modelle entwickelt. Eine Fülle an digitalen wissenschaftlichen Informationen ist somit im Internet frei zugänglich (vgl. Bargheer, Bellem & Schmidt, 2006, S. 3–7).

Das World Wide Web eignet sich hervorragend für die Präsentation, den Vertrieb und die Rezeption von digitalen Ressourcen aller Art. Hinsichtlich der Suche nach Informationen sind zwei Aspekte von Bedeutung: die Menge an Informationen und die Organisationsstrukturen der Dokumente im Web. Neben qualitativ hochwertigen Informationen finden sich im Internet auch zahlreiche Webseiten und Dokumente mit irrelevanten Inhalten. Das exponentielle Wachstum der Dokumentenmenge im Internet und die fehlenden internen Ordnungsstrukturen hat die Informationsanbieter dazu veranlasst, verschiedenartige Instrumente zur Präsentation und Bereitstellung von Informationen zu entwickeln (vgl. Rösch & Weisbrod, 2004, S. 177–188; Pieper & Wolf, 2009, S. 356–359).

Die modernen Informationstechnologien haben die Suche nach Informationen im Allgemeinen vereinfacht. Jedoch wird in der Literatur zu Recht auch von der „frustrating experience“ gesprochen (vgl. Chang, Healey, McHugh & Wang, 2001, S. XIII), die bei der Suche nach relevanten Informationen im Web entstehen kann und welche unter anderem in der Fülle an komplexen Möglichkeiten zum Suchen und Finden von Informationen begründet ist (vgl. Pieper & Wolf, 2009, S. 356–359). Es wurden Techniken und Dienstleistungen entwickelt, die einen effizienteren Zugriff auf wissenschaftsrelevante Literaturquellen im Internet ermöglichen, wie beispielsweise wissenschaftliche Suchmaschinen.

Chang et al. (2001, S. 3–4, S. 19–20, S. 35–38) unterscheiden grundsätzlich zwischen mehreren Typen von Suchmaschinen:

- Keyword-basierte Suchmaschinen, welche wiederum in Suchmaschinen (*Search Engines* oder *Web Indexes*), Web Directories, Meta-Suchmaschinen und Information-Filtering-Tools unterteilt werden.
- Abfrage-basierte Suchsysteme: für die, basierend auf der Standard-Abfragesprache Structural Query Language (SQL), weitere Webabfragesprachen mit komplexeren Suchfunktionalitäten entwickelt wurden, z.B. WebSQL.
- Mediators und Wrappers: Mediatoren sind Systeme, die auf den Webabfragesprachen basieren und Suchmaschinen im Hintergrund nutzen, um Informationen aus dem Web abzufragen. Mediatoren übernehmen eine Vermittlerrolle zwischen Nutzer (*Client*) und Server und haben die Aufgabe, eine Nutzeranfrage bedarfsgemäß umzuwandeln und an die Suchmaschine für die Weiterverarbeitung weiterzuleiten. Die Antworten der Suchmaschine werden von den Mediatoren umgewandelt, bevor diese an den Nutzer weitergegeben werden. Die Wrappers sind Systemkomponenten, die von den Mediatoren verwendet werden, um die Daten aus dem Web zu extrahieren. Wrappers haben die Aufgabe, Daten zu filtern und in benötigte Formate umzuwandeln.
- Multimedia-Suchmaschinen.

Bei der Auswahl der Internet-Suchdienste wurde bewusst das Hauptaugenmerk auf die Spezialsuchmaschinen gelegt, die zu der ersten der oben beschriebenen Kategorien gehören. Lewandowski (2009, S. 53–59) schildert zunächst einige Aspekte, denen die allgemeinen Suchmaschinen nicht zufriedenstellend entsprechen können, um anschließend die besonderen Suchmöglichkeiten der spezialisierten Suchmaschinen zu präsentieren und sie mit den Universalsuchmaschinen zu vergleichen. Lewandowski (2009, S. 57–58) bezeichnet diejenigen Suchmaschinen, die sich thematische oder formale Beschränkungen auferlegen, als Spezialsuchmaschinen und unterscheidet zwischen folgenden Typen:

- Spezialsuchmaschinen für bestimmte Bereiche des Web, die mittels „focused crawling“ Dokumente erfassen und dabei Listen von Webservern (*white lists*) für den Aufbau des Datenbestandes berücksichtigen. Der Vorteil dieser Arbeitsweise liegt darin, dass durch die Vorauswahl der Server bereits vorab

Qualitätskontrolle und thematischer Fokus erfolgen.

- Spezialsuchmaschinen für bestimmte Dokumententypen, die eine besondere Auswertung der spezifischen Eigenschaften des jeweiligen Dokumententyps ermöglichen.
- Spezialsuchmaschinen für aktualitätskritische Dokumente, die innerhalb von kurzen Aktualisierungsintervallen die zeitnahe Erfassung von bestimmten Dokumententypen fokussieren (z.B. Nachrichten, Blogs).
- Hybridsuchmaschinen verbinden die Erschließung von Dokumenten aus dem freien Web mit solchen aus dem Invisible Web, die sowohl kostenlose auch kostenpflichtige Inhalte enthalten können.

Für die vorliegende Aufgabenstellung wurden hauptsächlich Wissenschaftssuchmaschinen verwendet, die eine Art von Spezialsuchmaschinen darstellen.
Nach Lewandowski (2009, S. 62–63) indexieren die speziellen Wissenschaftssuchmaschinen entweder „versteckte“ Datenbankinhalte oder den wissenschaftlichen Teil des „sichtbaren“ Web, oder aber sie ermöglichen die Suche nach Inhalten aus „beiden Welten“ unter einer Oberfläche.
Im nächsten Kapitel werden beispielhaft einige der verwendeten Literaturdatenbanken und Suchmaschinen mit ihren wichtigsten Merkmalen dargestellt.

3.3 Kurzbeschreibung einiger Rechercheinstrumente

ENTEC Energietechnik

Die Datenbank ENTEC, ursprünglich in Kooperation zweier Forschungsinformationszentren (FIZ Karlsruhe[32] und FIZ Technik[33]) erstellt, wird seit 2006 von WTI Frankfurt e.G.[34] betreut. Entec ist eine bibliographische Datenbank und stellt einen Ausschnitt der Datenbank TEMA dar. Die Datenbank enthält ca. 570.000[35] Literaturnachweise und wird wöchentlich aktualisiert. Entec weist Literatur seit dem Erscheinungsjahr 1968 aus den Gebieten der Energietechnik und Energieforschung nach. Hierfür werden die einschlägigen nationalen und internationalen

[32] http://www.fiz-karlsruhe.de
[33] http://www.fiz-technik.de/tecfinder/faces/facelets/search/search.jsp
[34] http://www.wti-frankfurt.de/
[35] Stand: 07.01.2013

Publikationen (Zeitschriftenaufsätze, Konferenz- und Forschungsberichte, Dissertationen) ausgewertet. Die Literaturnachweise werden durch Abstracts beschrieben und mit englischen und deutschen Schlagwörtern aus dem *Thesaurus Technik und Management* erschlossen. Die Literaturnachweise werden durch die Zuweisung von Codes aus der Klassifikation *Fachordnung Technik* einem Fachgebiet zugeordnet.

Es werden drei Suchlevels angeboten: einfache, erweiterte und Strategiesuche. In der Strategiesuche können 20 Datenbankfelder durchsucht werden. Bei der Suche können die logischen Operatoren AND und OR verwendet werden. Trunkierung und Phrasensuche wird in bestimmten Suchfeldern ermöglicht. Die Datenbank ist von der Technischen Informationsbibliothek (TIB)[36] lizensiert und in DBIS verzeichnet.

ETDEWEB

Nach McKeating und MacLeod (2001, S. 289–292) ist dies die wichtigste Datenbank für den Bereich der Energietechnik. Sie wird vom amerikanischen Department of Energy koordiniert und enthält ca. 5 Millionen[37] Publikationsnachweise, ca. 500 Tausend Volltexte und über eine Million permanente Links zu Volltexten auf Verlagsservern. Die Literaturnachweise, die bis 1974 zurückreichen, werden von den Mitgliedern der International Energy Agency's Energy Technology Data Exchange (ETDE) und der International Energy Agency zusammengestellt.

Inhaltliche Schwerpunkte der Datenbank liegen in folgenden Bereichen: Energiespeicherung, -nutzung und -management, umweltbezogene Aspekte der Energienutzung, Energieressourcen, Solar-, Wind- Wasserenergie und andere alternative Energiequellen sowie Atomenergie und Werkstoffwissenschaften. Aus diesen Gebieten werden unterschiedliche Publikationstypen, wie z.B. wissenschaftliche, technische und Konferenzberichte, Fachbücher, Dissertationen und Patentschriften ausgewertet.

[36] www.tib.uni-hannover.de
[37] Stand: Februar 2013

Suchmöglichkeiten für unterschiedliche Niveaus werden angeboten. In der erweiterten Suche können Begriffe in unterschiedlichen Datenbankkategorien über mehrere Felder gesucht werden, die mit *Und* verknüpft werden. Hier ist eine Suche nach Autoren im entsprechenden Index möglich sowie eine Schlagwortsuche mit Hilfe von Deskriptoren aus dem *International Energy Subject Thesaurus* (vgl. McKeating & MacLeod, 2001, S. 289–292), die direkt aus dem Thesaurus in die Suche übernommen werden können. Die Literaturnachweise sind mit Schlagwörtern erschlossen und enthalten zusätzlich Inhaltsangaben in Form von Abstracts.

Für die Erstellung einer Suchstrategie dürfen die Booleschen Operatoren AND, OR, NOT und der Proximity Operator NEAR sowie Wildcards verwendet werden. Die Suche nach exakten Phrasen ist ebenso möglich wie die nach ungefähren Phrasen (*inexact phrase*). Ein kostenloser Zugriff auf die Datenbank ETDEWEB[38] wird vom FIZ Karlsruhe für alle deutschen Kunden angeboten. Die Datenbank ist im DBIS verzeichnet.

GetInfo

Das Portal für technisch-naturwissenschaftliche Fach- und Forschungsinformationen wird von der Technischen Informationsbibliothek in Hannover betreut. Das Portal ermöglicht eine Recherche in den Beständen der TIB, der Zentralbibliotheken für Medizin und Wirtschaftswissenschaften, in Fachdatenbanken sowie in Verlagsangeboten, insgesamt in mehr als 150 Millionen internen und externen Datenquellen. Neben bibliographischen Angaben werden Volltexte, Forschungsdaten, audiovisuelle Medien und 3D-Modelle nachgewiesen.
Sowohl die einfache als auch die erweiterte Suche ermöglichen die Nutzung von Booleschen Operatoren und den Einsatz von Trunkierung. Es stehen keine Indizes, Thesauri oder Klassifikationen für die Suche zur Verfügung. Die Literaturnachweise sind in GetInfo inhaltlich nicht erschlossen. In der erweiterten Suche kann nach Begriff, Titel, Autor, Jahr und nach Publikationsnummer (ISBN oder

[38] https://www.etde.org/etdeweb

ISSN) gesucht werden. Das Portal GetInfo[39] kann über DBIS oder über die Webseite der TIB aufgerufen werden.

INSPEC Information Services in Physics, Electronics and Computing

Die Datenbank INSPEC[40] wird von der Institution of Engineering and Technology (IET) produziert und ist eine der wichtigsten Datenbanken für unterschiedliche Bereiche der Ingenieurwissenschaften. Internationale Publikationen (Zeitschriften-, Kongressbeiträge, Fachbücher, Dissertationen und technische Berichte) aus den Gebieten der Elektronik und Elektrotechnik, Physik, Informatik und Informationstechnologien sowie aus verwandten Fächern werden mit bibliographischen Angaben und Abstracts nachgewiesen. In INSPEC ist die seit 1969 erschienene Literatur erfasst, die Datenbank enthält über 13 Millionen[41] Datensätze und wird wöchentlich aktualisiert.

Alle Datensätze sind indexiert und durch Inhaltsangaben in englischer Sprache erschlossen. Da der Zugang zu INSPEC über WTI Frankfurt e.G. erfolgt, stehen die gleichen Suchfunktionen wie bei der Datenbank ENTEC zur Verfügung. In Inspec wird die Möglichkeit angeboten, mit kontrolliertem Vokabular aus dem *Inspec Thesaurus* zu suchen. Ein Index mit den Deskriptoren wird unter der Option *Thesaurussuche* angeboten. Der Inspec-Thesaurus enthält ca. 19 Tausend Termini[42] mit 9722 Vorzugstermini und besteht aus zwei Sektionen: aus einer alphabetischen und einer hierarchisch aufgebauten Liste von Termini. Im hierarchischen Teil des Thesaurus werden die Beziehungen zwischen den Termini, die verwendeten Synonyme, das Datum der Aufnahme und die Termini, welche vor der Aufnahme verwendet wurden, verzeichnet (vgl. McKeating & MacLeod, 2001, S. 287–289). Die erfasste Literatur wird zusätzlich durch freie Schlagwörter erschlossen und mittels Codes aus der *INSPEC Classification List*[43] den entsprechenden Fachgebieten zugeordnet. Die Datenbank INSPEC ist von der TIB lizenziert und in DBIS nachgewiesen.

---

[39] http://www.tib-hannover.de/de/getinfo/
[40] http://www.theiet.org/resources/inspec/
[41] Stand: 16.01.2013
[42] Stand: 2012
[43] http://www.theiet.org/resources/inspec/about/records/iclass.cfm

SciVerse HUB

Die Inhalte der Elsevier-Produkte Scopus[44] und ScienceDirect[45] können über eine gemeinsame Oberfläche, SciVerse HUB[46], recherchiert werden. Scopus wurde von Elsevier 2004 auf den Markt gebracht und galt schon 2008 als eine der umfangreichsten Literaturdatenbanken weltweit (vgl. Sühl-Strohmenger, 2008, S. 24, S. 202). Inzwischen enthält SciVerse über 500 Millionen Datensätze: Zeitschriftenaufsätze, Fachbücher, Dissertationen, Patente und über 500 Millionen wissenschaftliche Webseiten aus den Gebieten der Natur- und Sozialwissenschaften, Technik und Medizin werden nachgewiesen. Über ScienceDirect werden mehr als neun Millionen Volltexte aus zahlreichen Büchern und Fachzeitschriften zur Verfügung gestellt. Scopus enthält lediglich bibliographische Angaben und Abstracts.

Die Recherche kann über SciVerse HUB in beiden Produkten und im wissenschaftlichen Web gleichzeitig oder aber nur in einem Teil der Datenbank erfolgen. Zum Aufbau der Suchanfragen können die Booleschen Operatoren, die Proximity Operatoren WITHIN und NEAR, Wildcards für die Suche nach unterschiedlichen Wortformen sowie die Phrasensuche verwendet werden.

Browsen nach Titeln oder nach Themengebieten ist ebenso möglich wie die Suche mit Begriffen. Die Feldersuche ist in beiden Produkten und unter der gemeinsamen Suchoberfläche der gesamten Datenbank möglich. In SciVerse HUB kann z.B. nach Autor, Titel, Dokumenttyp, Thema, Jahr und Konferenzname gesucht werden. Dabei muss beachtet werden, dass die beiden Teildatenbanken spezifische Suchfeldbezeichnungen haben. Neben einer erweiterten Suche, in der die Suchfeldbezeichnung ausgewählt werden kann, wird auch eine Expertensuche angeboten, in der komplexe Suchanfragen mittels einer spezifischen Kommandosprache entwickelt werden können. SciVerse HUB ist für die Recherche frei zugänglich im Web und wird im DBIS verzeichnet. Die Zugänglichkeit der Rechercheergebnisse ist jedoch lizenzpflichtig.

[44] http://www.info.sciverse.com/scopus
[45] http://www.info.sciverse.com/sciencedirect
[46] http://www.hub.sciverse.com/action/home

SpringerLINK

Die Online-Plattform des Springer Verlags mit der Bezeichnung SpringerLINK[47] besteht seit 1996 und umfasst seit der Fusion mit Kluwer Academic Publishers in 2004 ein umfangreiches Publikationsspektrum beider Verlage (vgl. Schnaub, 2005, S. 90). Die Datenbank enthält Literaturnachweise mit bibliographischen Angaben und Abstracts aus elektronischen Zeitschriften, E-Books, Schriftenreihen, technischen Berichten sowie Nachschlagewerken. Die digitalen Inhalte des interdisziplinären Angebots werden thematisch in 25 *Online Libraries* organisiert, die für die entsprechenden Fachgebiete stehen. Auf die Volltexte kann bei vorhandener Lizenz zugegriffen werden. Artikel mit der Kennzeichnung *Open Access* stellt Springer kostenfrei im Volltext zur Verfügung.

Für die Recherche im gesamten Angebot stehen eine einfache Suche mit umfangreichen Filteroptionen sowie eine erweiterte Suche zur Verfügung, welche mehrere Suchfelder für die Erstellung einer präzisen Suchanfrage aufweist. Für den Aufbau von komplexen Suchanfragen können Boolesche und Proximity Operatoren sowie Wildcards verwendet werden. Aufgrund der Lemmatisierung der in der Datenbank vorhandenen Wortformen findet eine umfassende Suche statt. Das Ergebnis einer Suchanfrage enthält alle Treffer, die den Wortstamm enthalten. Das gilt auch für die Phrasensuche. Auf die in einem Beitrag zitierten Literaturquellen wird mittels Cross-Reference-Linking verwiesen. Die TIB hat die Sammlungen des Springer-Verlags lizenziert. SpringerLink ist in DBIS nachgewiesen.

3.4 Zusammenfassung

Es wurden beispielhaft Informationsquellen ausgewählt, die stellvertretend für den jeweiligen Quellentyp stehen. Die Datenbanken bieten qualitativ hochwertige Informationen zu einem oder mehreren Fachgebieten sowie viele Funktionen, die ein effizientes Recherchieren ermöglichen. Die Inhalte werden von Experten geprüft, ausgewählt, indexiert und strukturiert. Dadurch ist die Recherche zu speziellen Fragestellungen mittels komplexer Suchanfragen möglich, aber auch umgekehrt

[47] http://link.springer.com/

kann eine komplexe Suchanfrage verallgemeinert oder verfeinert werden. Die Ergebnisse sind fachlich relevante Treffer mit hochwertigen bibliographischen Daten, die nach verschiedenen Kriterien sortiert werden können. Somit sind die Datenbanken unerlässliche Informationsquellen für die professionelle Fachrecherche.

Durch die ständige Weiterentwicklung der Suchmaschinentechnologien bieten die Spezialsuchmaschinen und Fachportale auch für professionelle Informationsspezialisten gute Möglichkeiten, komplexe Rechercheanfragen umzusetzen und schnell relevante Literaturnachweise zu erhalten. Für das domainspezifische Training der automatischen Spracherkennung ist die Recherche mittels wissenschaftlicher Suchmaschinen unerlässlich. Die Recherche in den freien Internetquellen führt unvermeidlich zu Dubletten. Die Prüfung auf und die eventuelle Eliminierung von Dubletten bedeutet für den Rechercheur hohen Zeitaufwand. Somit können die speziellen Internet-Suchdienste nur eine Ergänzung zu den professionellen Fachdatenbanken darstellen.

4. Adaption des Sprachmodells der automatischen Spracherkennung

Automatische Spracherkennungssysteme (Automatic Speech Recognition - ASR) können während des Spracherkennungsprozesses nicht alle Wörter korrekt erkennen und daher noch keine guten Transkriptionen erstellen. Es gibt viele Faktoren, die die Qualität der automatischen Spracherkennung beeinflussen, z.B. Größe des Vokabulars, diskontinuierliche, spontane Rede oder Gespräche, Umweltgeräusche, sprecherspezifische Besonderheiten und die Eignung des Mikrofons (vgl. Jurafsky & Martin, 2009, S. 320–321).

Die Sprachmodelle von ASR sind mit Hilfe von Korpora zu allgemeinen Themen trainiert. Charakteristisch für die Sprachmodelle ist eine Funktionsweise, die auf Berechnung der Wahrscheinlichkeiten von Wortsegmenten basiert. Es handelt sich dabei um die sogenannten N-gram-Wahrscheinlichkeiten, die gesammelt und global für ein großes Korpus berechnet werden. N-gram-Modelle verwenden die letzten N-1-Wörter, um die Wahrscheinlichkeit für das nächstmögliche Wort zu berechnen. Das Sprachmodell unterstützt ASR bei der Auswahl der Wortseg-

mente mit der höchsten Wahrscheinlichkeit. Die Abbildung 4-1 zeigt den Aufbau eines Spracherkenners.

Allgemeine Sprachmodelle sind für die Transkription von gesprochenen Dokumenten, in denen ein oder mehrere fachspezifische Themen sukzessive behandelt werden, wenig geeignet (vgl. Lecorvé, Gravier & Sébillot, 2008, S. 592; Mahajan, Beeferman & Huang, 1999, S. 541). Je vielfältiger und komplexer die Themen, desto größer die Anzahl der fachspezifischen Wörter ist, die erkannt werden müssen, desto schwieriger die Erkennungsaufgabe und desto schlechter die Transkriptionsergebnisse. Das Basis-Sprachmodell soll deshalb in einem iterativen Prozess durch den Aufbau von themenspezifischen Sprachmodellen modifiziert werden.

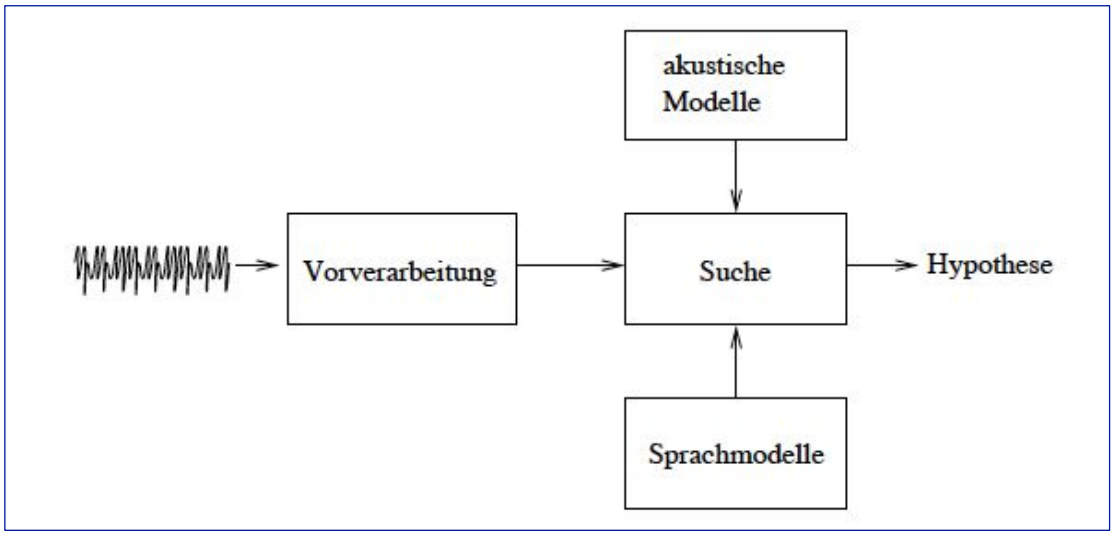


Abb. 4-1 Aufbau eines Spracherkenners. Quelle: Schaaf, 2004, S. 25

Die automatische Transkription von Vorlesungsreden ist immer noch eine Herausforderung für die ASR-Systeme und stellt einen der meistuntersuchten Bereiche dar. Ziel der Forschung ist die Entwicklung von separaten Sprachmodellen, welche die Besonderheiten von Vorlesungsreden berücksichtigen: ein umfangreiches Fachvokabular, kontinuierliche Sprache, jedoch mit vielen Unterbrechungen, geschuldet der Spontaneität von Vorlesungsreden. Nach Munteanu (2009, S. 71) hat das ideale Sprachmodell ein großes Vokabular, ist sprecherunabhängig sowie themen- und domänenspezifisch.

In diesem Kapitel werden die unterschiedlichen Techniken zur Adaption von Sprachmodellen dargestellt und die Rahmenbedingungen sowie das verwendete Verfahren bei der in dieser Arbeit durchgeführten Untersuchung (Abschnitt 4.1). Die für das Training des Sprachmodells und das Testen der Adaption verwendeten Datentypen, deren Umfang und die darin behandelte Thematik werden im Abschnitt 4.2 beschrieben. Daran anknüpfend werden im Abschnitt 4.3 die damit durchgeführten Experimente und deren Ergebnisse beschrieben. Im Abschnitt 4.3 werden die Ergebnisse zusammengefasst.

4.1 Themenspezifische Adaption des Sprachmodells

Ziel der Adaption von Sprachmodellen ist die Veränderung der Parameter des Sprachmodells, damit es auch bei Gebieten mit speziellem Vokabular gute Resultate erzielen kann. Nach Fügen (2009, S. 67–69) ist die Interpolation von Modellen eine der wichtigsten Techniken zur Adaption des Sprachmodells von automatischen Spracherkennungssystemen. Dabei werden mehrere Modelle linear ineinander gefügt, um ein gemischtes Modell zu formen. Die Adaption eines Sprachmodells kann durch die Aktualisierung von Interpolationsgewichtungen oder durch Austausch oder Hinzufügung von neuen Modellkomponenten erfolgen, die zum Thema besser passen. Die neuen Modellkomponenten können nach Fügen (2009, S. 67) sprecher-, rede- oder themenspezifisch modelliert werden.

Hinsichtlich der Adaption von Sprachmodellen wird unterschieden zwischen: *within-domain* und *cross-domain* Adaption. Bei der *within-domain* Adaption werden Sprachmodelle für unterschiedliche Themen in einer einzigen Domäne angepasst. Diese Art von Adaption bezweckt, die Parameter des Sprachmodells durch den Einsatz von Trainingsdaten feiner zu justieren, bis sich die Leistungen des Sprachmodells für die gewünschten Themen deutlich verbessert haben. Bei der *cross-domain* Adaption wird ein Sprachmodell an ein neues angepasst, für das nur eine begrenzte Anzahl von Trainingsdaten zur Verfügung steht. Deshalb wird bei dieser Art von Adaption von einer Hintergrund-Domäne gesprochen, die dem Basis-Sprachmodell zugrunde liegt. Die neue Domäne mit der kleinen Menge von Trai-

ningsdaten wird als Adaptionsdomäne bezeichnet (vgl. Suzuki und Gao, 2005, S. 265–272).

Zwei Hauptmethoden werden bei der Durchführung von cross-domain Adaptionen verwendet: *maximum a posteriori* (MAP) Schätzungsmethoden und auswählende (*discriminative)* Trainingsmethoden. Bei der MAP-Methode werden die neuen Adaptionsdaten dazu verwendet, eine Anpassung von Parametern des Basismodells zu erreichen. Die von Fügen (2009, S. 67–69) erwähnte Interpolation von zwei Sprachmodellen ist eine MAP-Methode. Die auswählenden Methoden verringern durch die Nutzung von Adaptionsdaten unmittelbar die Fehler, die das Basismodell aufgrund von Adaptionsdaten gemacht hat (vgl. Suzuki & Gao, 2005, S. 265–272).

Kneser und Peters (1997, S. 779–782) entwickelten ein auf semantischen Klassen basierendes Verfahren zur Adaption von Sprachmodellen, welches die Verteilung von Kookkurrenzen im Korpus berücksichtigt. Das Verfahren geht von der Annahme aus, dass in einem Text in der Regel mehrere Themen behandelt werden. Ihr Modell der semantischen Klassen sieht vor, dass jedes Wort im Wörterbuch der Spracherkennung genau einer Klasse zugeordnet werden kann und sich nur die Themenanteile innerhalb eines Textes verändern, nicht die Klassenzugehörigkeit der Wörter. Das ist die Grundlage für den von ihnen entwickelten Algorithmus, welcher die Aufgabe hat, maximale Ähnlichkeiten zu berechnen, mittels relativer Wahrscheinlichkeiten auf der Basis der gezählten Kookkurrenzen innerhalb der Trainingsdaten. Dadurch werden die Wörter zu Clustern gruppiert, die den semantischen Klassen entsprechen. Basierend auf diesen semantischen Klassen wurden angepasste unigram, bigram und trigram Komponenten entwickelt.

Mahajan, Beeferman und Huang (1999, S. 541) erklären ebenfalls, wie domänen- oder themenspezifische Sprachmodelle durch automatisches Clustering die Trainingsdaten nach Gruppen von ähnlichen Themen trennen, so dass einzelne Segmente von Trainingsdaten unterschiedlichen Themen zugeordnet werden können. Die Autoren weisen auf die Folgen der Kombination von unterschiedlichen themenspezifischen Trainingsdaten und auf den stets steigenden Schwie-

rigkeitsgrad bei der Berechnung von Wahrscheinlichkeiten aller möglichen Wortsegmentkombinationen hin. Sie empfehlen deshalb ein Verfahren zur Entwicklung eines dynamischen Sprachmodells, das weder vordefiniertes Clustering noch eine Segmentierung von Trainingsdaten erfordert (vgl. Mahajan, Beeferman & Huang, 1999, S. 542). Sie nutzen die Dokumenthistorie für die Formulierung einer Datenabfrage und eine Information-Retrieval-Technik, die TF-IDF (vgl. Kap. 1.1 bzw. 5.1.1), um ähnliche Dokumente innerhalb der gesamten Trainingsdaten zu identifizieren und diese nach Relevanz zu selektieren. Die relevantesten Dokumente bilden die Sammlung für die Adaption eines themenunabhängigen Sprachmodells. Die Vorteile von *Multiple Dynamic Topic Language Models*, wie die Entwickler sie nennen (vgl. Mahajan, Beeferman & Huang, 1999, S. 543), liegen in der Anpassungsfähigkeit an neue Themen des themenunabhängigen und gleichzeitig themenspezifischen Sprachmodells.

Munteanu (2009, S. 82–87) entwickelte ein Verfahren für die web-basierte Adaption des Sprachmodells für die automatische Transkription von Vorlesungen, bei dem weder der komplexe Prozess der Ermittlung und Erwerbung von bestpassenden Korpora für den Aufbau mehrerer spezifischer Sprachmodelle noch die Extraktion von Schlüsselwörtern aus zusätzlichen Informationsquellen notwendig ist. Er nutzt den gesamten Inhalt der Folien in der Präsentation, um Abfragen für die Websuchmaschinen zu formulieren und verwendet die auf diese Weise gesammelten Webkorpora, um ein einziges Sprachmodell zu trainieren, das den Ansprüchen spontaner Rede gerecht wird und zugleich besser zu den Besonderheiten von Vorlesungen passt.

4.1.1 Rahmenbedingungen für die Adaption

Das Kompetenzzentrum für nicht-textuelle Materialien[48] (KNM) nutzt für die Audioanalyse die Software für automatische Spracherkennung der Firma European Media Laboratory (EML)[49]. Das installierte Vokabular hat einen Umfang von ca. eine Million Wortformen. Die Spracherkennung unterstützt Deutsch und Englisch.

Ausgangspunkt für das Training von themenspezifischen Sprachmodellen bildete die durch die TIB durchgeführte Evaluation von Transkriptionsergebnissen im März 2012[50]. Die Fehlerrate der nicht korrekt erkannten Wörter *(Word Error Rate - WER*) lag bei den bewerteten Transkriptionen zwischen 32% und 64%. In der Abbildung 4-2 werden die prozentualen Werte von WER für fünf Videotranskriptionen veranschaulicht. Die blaue Linie zeigt die von der EML erhobenen Evaluationsergebnisse. Die grüne Linie zeigt die Fehlerraten, die bei der Evaluation durch die TIB mit einem Programm[51] des Hasso-Plattner-Instituts[52] gemessen wurden. Zwei weitere Werte werden mit den roten und gelben Kurven dargestellt und beziehen sich auf Humanevaluation. Diese Aspekte wurden in der vorliegenden Arbeit nicht behandelt. (vgl. dazu auch Anlage C1).

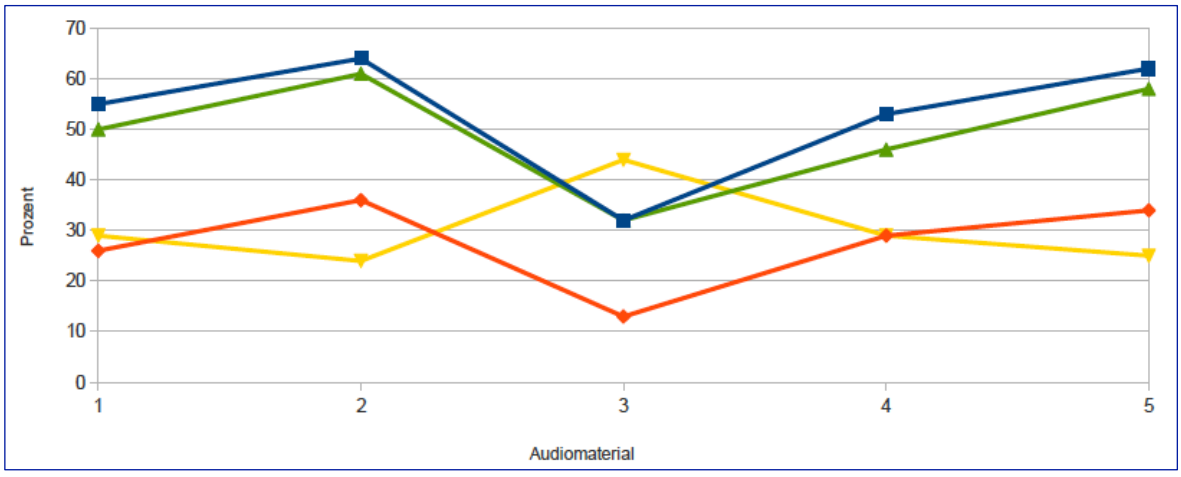


Abb. 4-2: Ursprüngliche Word Error Rate bei der Evaluation von Transkriptionsergebnissen

[48] http://www.tib-hannover.de/de/dienstleistungen/kompetenzzentrum-fuer-nicht-textuelle-materialien-knm/
[49] http://www.eml-development.de/index_d.php
[50] Die Angaben sind einem mir zur Verfügung gestellten internen Dokument entnommen.
[51] Die Bezeichnung des Programms ist der Verfasserin nicht bekannt.
[52] http://www.hpi.uni-potsdam.de/willkommen.html

4.1.2 Verfahrensweise bei der Adaption des Sprachmodells

Unter Berücksichtigung der Vorgaben des Kompetenzzentrums für nicht-textuelle Materialien und der Anforderungen des Systems, mit dem die Adaption durchgeführt werden sollte, wurde in Anlehnung an Fügen (2009, S. 67–97) folgendes Verfahren für die themenspezifische Adaption des Sprachmodells entwickelt:

- Zuerst wurde eine Recherche in GetInfo[53] nach Videos aus einem Fachgebiet durchgeführt und die ausgesucht, bei denen Titel und die Namen der Sprecher ersichtlich waren.
- Das Thema der Rede / der Vorlesung wurde aus der Inhaltsbeschreibung oder aus dem Inhalt des Videos selbst ermittelt.
- Diese Angaben bildeten die Grundlage für die Ableitung von Suchbegriffen.
- Die festgelegten Suchbegriffe wurden für die Formulierung von Suchanfragen verwendet. Es folgte die Entwicklung einer Suchstrategie. Die Prinzipien hierfür wurden im Kapitel 3.1 dargestellt.
- Für die Umsetzung der Suchstrategie wurden die Rechercheinstrumente ausgewählt und die Recherche manuell durchgeführt.
- Jedes Textdokument und jede Webressource wurde vor der Aufnahme in die Trainingsdaten zuerst auf Eignung geprüft.
- Nach dem Aufbau einer Datensammlung für das Training des Sprachmodells wurde das Training auf der dafür verwendeten Plattform angestoßen. Das System übernimmt dann das Filtern von Daten (z.B. Entfernen von HTML-Tags, Bildern, Metadaten, Menüs), die Satz- und Wortsegmentierung, die erforderliche Normalisierung vor der Analyse und die Interpolation des neuen mit dem vorhandenen Sprachmodell.
- Um die Ergebnisse des Trainings zu testen, wurde analog zu den Trainingsdaten nach Testdaten im Internet recherchiert.
- Nach der erforderlichen Vorverarbeitung wurden die Testdaten auf der Trainingsplattform hochgeladen und das Testen mit dem Basis-Sprachmodell angestoßen. Abschließend wurden die Ergebnisse der Adaption und der Tests mit dem Basis-Sprachmodell evaluiert.

In der Abbildung 4-3 wird das dargestellte Verfahren grafisch veranschaulicht.

[53] https://getinfo.de/app

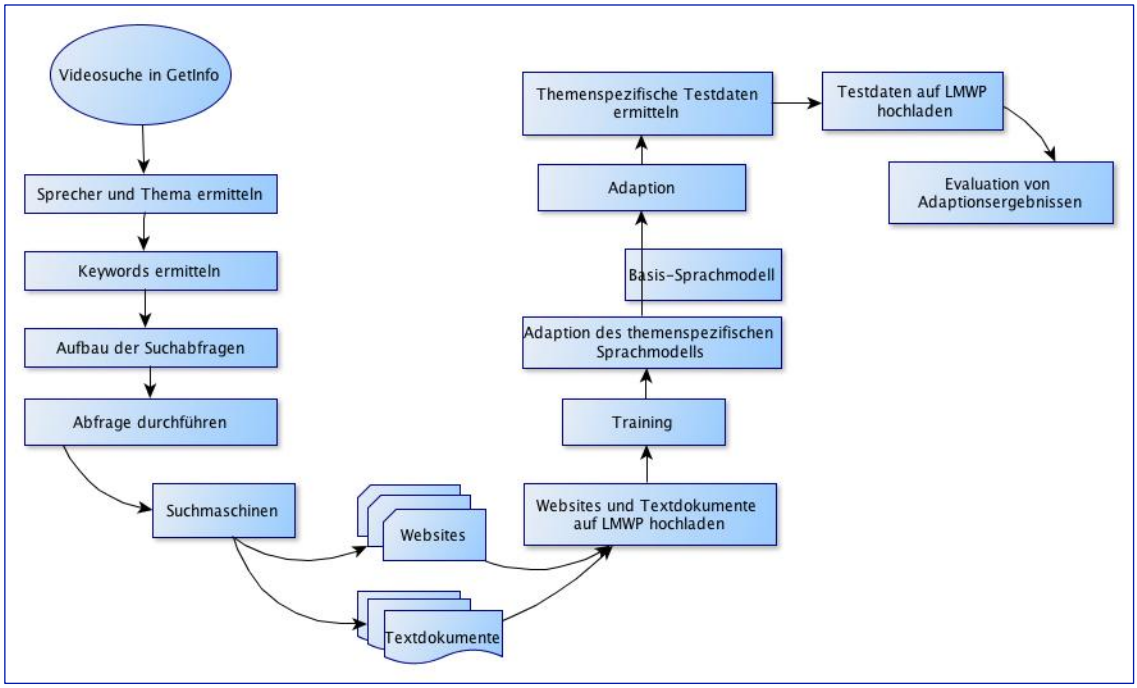


Abb. 4-3: Verfahren der themenspezifischen, web-basierten Adaption des Sprachmodells

4.1.3 Verwendetes System für die Adaption des Sprachmodells

Die Adaption des Basis-Sprachmodells erfolgte mit der webbasierten Komponente *Language Model Workplace* (LM Workplace) der Transkriptionsplattform der European Media Laboratory GmbH (EML). Die integrierten Tools ermöglichen es, themenspezifische Sprachmodelle durch das Hochladen von Webdokumenten oder Textsammlungen zu entwickeln und das Basis-Sprachmodell automatisch anzupassen. Das Filtern der Daten aus den Webdokumenten und der gesamte Adaptionsprozess des Sprachmodells sind vollständig automatisiert. Das System ermöglicht es, für die themenbasierte Adaption einzelne Bereiche einzurichten, nach Fachgebiet und Sprache unterschieden. Die Auswahl des richtigen Basis-Sprachmodells (*Language Model - LM*) und des akustischen Modells (*Acoustical Model - AM*) erfolgt automatisch. Der Nutzer kann den Stand des Adaptions-prozesses jederzeit einsehen, ebenso eine chronologische Liste mit allen durchgeführten Adaptionsphasen und die Ergebnisse der Adaption, sowohl des neuen, spezifischen als auch des Basis-Sprachmodells. In der Abbildung 4-4 wird die Architektur von LM Workplace veranschaulicht.

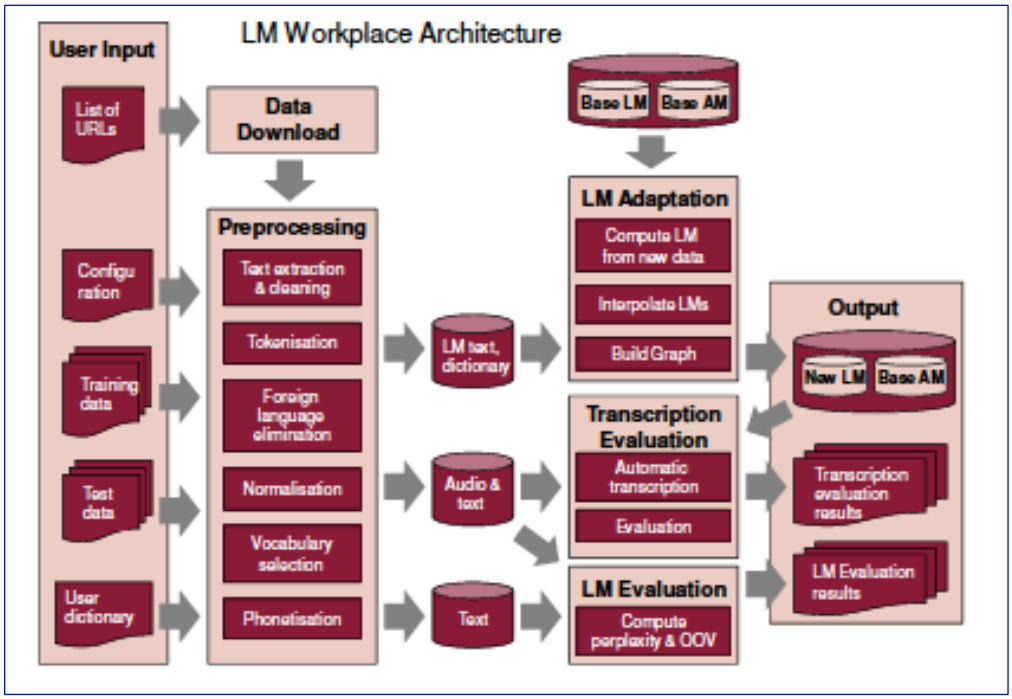


Abb. 4-4: Architektur von Language Model Workplace.
Quelle: LM Workplace User guide, 2012

4.2 Verwendete Trainings- und Testdaten

Fügen (2009, S. 69) empfiehlt für die Adaption eines Sprachmodells Transkriptionen gesprochener Sprache. Die gesprochene unterscheidet sich von der geschriebenen Sprache durch Unterbrechungen und Wiederholungen. Dies beeinflusst die Ergebnisse der automatischen Spracherkennung negativ. Es wurde deshalb Wert darauf gelegt, für das Training des Sprachmodells neben thematisch relevanten Text- und HTML-Dokumenten auch Transkriptionen zu identifizieren und zu verwenden.

Für die Adaption des englischen Sprachmodells wurde ein Training für die Fächer Informatik und Technik in mehreren Etappen durchgeführt. Es wurden hierfür drei Typen von Daten verwendet: Textdokumente, Transkriptionen von Reden und Workshop-Beiträgen sowie HTML-Dokumente (Webseiten und Forenbeiträge). Um relevante Dokumente zu den Themen Energietechnik, Robotik, Software-Programmierung, Internet, Datenanalyse und Datensicherheit nachzuweisen und zu selektieren, wurden hauptsächlich Websuchmaschinen eingesetzt. Verwendete Verfahren und Quellen hierfür wurden im Kapitel 3 dargestellt.

Für das Fach Informatik wurde das Modell mit insgesamt 1.356.538 Wörtern trainiert, die aus 64 Webseiten, 12 Textdokumenten und sechs Transkriptionen stammen. Für das Training des Faches Technik wurden 46 Webseiten, 6 Textdokumente und 12 Transkriptionen mit insgesamt 1.376.957 Wörtern verwendet.
Die Testdaten für das Fach Informatik umfassen 429 Minuten. Eine Transkription im Textformat und 9 Audiodateien mit den dazugehörigen Transkriptionen wurden zum Testen verwendet. Die Testdaten für dieses Fach hatten einen Umfang von 130.378 Wortformen. Mit 414 Minuten wurde die Adaption im Fach Technik getestet. Insgesamt umfassten die Testdaten 60.793 Wortformen, die aus 13 Transkriptionen und 12 dazugehörigen Audiodateien stammen. In der Tabelle 4-1 werden die Trainings- und Testdaten für die beiden Fächer gegenübergestellt. In der Anlage C2 befindet sich eine Darstellung der Trainings- und Testdaten.

Projekt-bezeichnung	**Trainingsdaten**			
	Anzahl Web-seiten	Texte	Transkriptionen	Anzahl Wörter
Informatik	64	12	6	1.356.538
Technik	46	6	12	1.376.957

Projekt-bezeichnung	**Testdaten**			
	Audiodatei + Transkription	nur Trans-kript	Anzahl Minuten	Anzahl Wörter
Informatik	9	1	429	130.378
Technik	12	1	414	60.793

Tab. 4-1: Trainings- und Testdaten für die Adaption des Sprachmodells

Die Trainingsdaten für das Fach Informatik wurden mit Hilfe von Suchmaschinen aus unterschiedlichen Quellen gesammelt und sind heterogen. Neben Webseiten und Textdokumenten wurden jeweils zwei Transkriptionen von universitären Vorlesungen, Podcast- und Konferenzbeiträgen verwendet. Die Vorlesungstranskriptionen der *Harvard Extension School*[54] behandeln die Themen *Internet* und *Datensicherheit.* Die Konferenzbeiträge stammen von der *Falling Walls International Conference on Future Breakthroughs in Science and Society*[55] und behandeln die Themen *Computer* und *Datenanalyse* in einem technisch-interdisziplinären Kontext. Die Podcasts behandeln das Thema der *Datenanalyse* im Kontext von *Big Data* und *Cloud Computing.* Im Durchschnitt dauern die Podcast- und Konferenzbeiträge 14 und die Vorlesungen 105 Minuten. Eine tabellarische Übersicht

[54] http://www.extension.harvard.edu/
[55] http://falling-walls.com/lectures

mit den verwendeten Transkriptionen für das Training des Sprachmodells befindet sich in der Anlage C3.

Die Testdaten für Informatik bestehen aus Vorlesungsaufnahmen an der Stanford University - Stanford Center for Professional Development[56], als Teil der Vorlesungsreihe *Introduction to computer science - Programming Methodology.* Die Kurse gehörten zu einführenden Veranstaltungen in *Software Engineering* und *Programmierung*. Eine Vorlesung stammt von der Princeton University[57], ist Teil der *Princeton University President's Lecture Series* und behandelt das Thema *Digitale Medien* im sozio-politischen und technologischen Kontext. Mit Ausnahme der Vorlesung der Princeton University, mit drei Sprechern, haben alle Vorlesungen nur einen Sprecher. Die Vorlesungen dauern durchschnittlich 50 Minuten. Nur bei der Transkription der Vorlesung der Princeton University wurde der Name des Erstellers angegeben. Es handelt sich dabei um eine manuell erstellte Transkription. Über die Erstellungsweise der anderen Transkriptionen können keine Aussagen gemacht werden. Eine tabellarische Übersicht mit den verwendeten Testdaten befindet sich in der Anlage C4.

Die Trainingsdaten für das Fach Technik enthalten, ebenso wie die für das Fach Informatik, neben Webseiten und Textdokumenten auch Transkriptionen. Von der California Energy Commission[58] stammen 12 Transkriptionen von Workshop-Diskussionen, in denen unterschiedliche Aspekte aus dem Themengebiet *Energietechnik* behandelt werden. In der Anlage C5 werden sie dokumentiert.

Die Testdaten für das Fach Technik bestehen aus sieben Reden und fünf Videovorlesungen. Die Reden sind Teil der Reihe *Ideas worth spreading from the TED Conference* von TED Technology, Entertainment, Design[59] und behandeln Aspekte von *alternativen Energiequellen*. Die Vorlesungsaufnahmen vom Massachusetts Institute of Technology[60] sind Teile der Vorlesungsreihen *Classical*

[56] http://scpd.stanford.edu/publicViewHome.do?method=load
[57] http://www.cs.princeton.edu
[58] http://www.energy.ca.gov/2011_energypolicy/documents/#11142011
[59] http://www.ted.com/pages/about
[60] http://ocw.mit.edu/index.htm

Mechanics, Principles of chemical science und *Aircraft Systems Engineering.* In den Kursen wurden unterschiedliche Aspekte zu *Robotik* und *Energie* behandelt.
Die Reden dauern durchschnittlich 16 Minuten, die Vorlesungen 48 Minuten. Die längste Vorlesung dauert 112, die längste Rede 28 Minuten, die kürzeste 4 Minuten. Der Schwierigkeitsgrad der Reden ist gering, verglichen mit dem der Vorlesungen des Massachusetts Institute of Technology. Eine tabellarische Übersicht mit den verwendeten Testdaten befindet sich in der Anlage C6-C6a.

4.3 Experimente zur Adaption des Sprachmodells

Ziel der Untersuchung war es herauszufinden, welche Anzahl fachspezifischer Daten benötigt wird, um den Prozentsatz der falsch erkannten Wörter und zugleich die Perplexitätswerte zu senken. Im folgenden Abschnitt werden drei automatisch generierte Evaluierungsmaße[61] beschrieben, die hier ausgewertet wurden, um die Wirkung der Adaption auf das Sprachmodell zu beurteilen.

Perplexität (*Perplexity - PP*)
Die Perplexität ist ein Maß für die Beurteilung der Leistung von statistischen Sprachmodellen, mit dem die mittlere Vorhersagegenauigkeit eines Modells gemessen wird. Nach Schaaf (2004, S. 35) wird bei der Berechnung der Perplexität „der Kehrwert des geometrischen Mittels über die auf einem Trainingskorpus bestimmten Wahrscheinlichkeiten $\mathrm{p}(w\,i|w\,1,\ldots w\,i-1)$ auf einem neuen ungesehenen Textkorpus bestimmt“. Das ist für die Testdaten wichtig, da diese Wörter oder Wortsegmente enthalten können, die im Trainingskorpus nicht registriert wurden. Damit beim Auftreten dieser unbekannten Wortformen die Perplexitätswerte nicht drastisch ansteigen, wird ihnen im Trainingskorpus eine positive Wahrscheinlichkeit zugewiesen. Die Perplexitätswerte können bei Verwendung eines konstanten Testkorpus Aussagen über die Eignung eines Sprachmodells für das Korpus machen. Niedrigere Perplexitätswerte stehen nicht immer im Zusammenhang mit einer Reduzierung der Wortfehlerrate. Sie zeigen aber, dass das Sprachmodell im Mittel bessere Vorhersagen macht. Der Spracherkenner hat

[61] Weitere Evaluierungsmaße, z.B. Echtzeit-Faktor (*Real Time Factor - RTF*), BLEU Skala Score, NIST Score, die jedoch keinen direkten Bezug zu dieser Arbeit haben, werden von Fügen (2009, S. 36–37) beschrieben.

einen geringeren Auswahlaufwand und das Ergebnis wird schneller gefunden (vgl. Schaaf, 2004, S. 35–36).

Wortfehlerrate (*Word Error Rate - WER*)
Die Wortfehlerrate ist ein Standardmaß für die Evaluation der Qualität von Spracherkennungssystemen. Dieses Evaluationsmaß drückt die Spanne der minimalen Korrekturen zwischen einer gegebenen Hypothese und ihrer Umsetzung in der Transkription aus. Die WER wird mittels folgender Formel berechnet:

$$\text{WER} = \frac{S+D+I}{N}$$

Hierfür wird Summe von Ergänzungen (*substitutions - S*), Löschungen (*deletions – D*) und Hinzufügungen (*insertions – I*) von Wörtern, die notwendig waren, um die Hypothese in Transkription umzusetzen, durch die Wörteranzahl (*N*) der Transkription geteilt (vgl. Fügen, 2009, S. 36).

Anteil der nicht vorhandenen Wörter im Vokabular (*Out of Vocabulary-Rate*)
Aufgrund der Häufigkeit von Wortformen, die in einem für ein bestimmtes Fachgebiet geeigneten Korpus vorkommen, wird das Vokabular für den Spracherkenner festgelegt. Diese Wortformen werden *Vokabular-Wörter* (*In Vocabulary*) genannt. Wörter, die im Korpus nicht vorkommen oder die geforderte Mindesthäufigkeit nicht erreicht haben, werden OOV-Wörter (*Out of Vocabulary*) bezeichnet. Der prozentuale Anteil der Wörter, die in einem gegebenen Korpus vorkommen, aber nicht im Wörterbuch des Spracherkenners, wird als OOV-Rate bezeichnet (vgl. Fügen, 2009, S. 36; Schaaf, 2004, S. 22).

Die Experimente wurden mit den im vorigen Kapitel beschriebenen Trainings- und Testdaten durchgeführt. Für alle Experimente wurde das im Kapitel 4.2 dargestellte Verfahren angewandt. Hierfür wurde innerhalb von zwei Fächern das Training des Sprachmodells für vier Themen mit insgesamt 110 Webseiten und jeweils 18 Textdokumenten und Transkriptionen durchgeführt. Der Gesamtumfang der Trainingsdaten beträgt 2.733.495 Wörter. Zum Testen der Adaptionswirkung wurden 843 Minuten Datenmaterial verwendet. Insgesamt wurden mit dem Basis-Sprachmodell zwei Transkriptionen und 20 Test-Sets, bestehend aus jeweils einer Audiodatei und der dazugehörigen Transkription, getestet. Der Umfang der zum Testen

verwendeten Transkriptionen beträgt 191.171 Wörter. In der Tabelle 4-2 werden die Gesamtzahlen zusammengestellt.

Trainingsdaten			
Anzahl Webseiten	Text-dokumente	Transkriptionen	Anzahl Wörter
110	18	18	2.733.495
Testdaten			
Audiodatei + Transkription	nur Transkription	Anzahl Minuten	Anzahl Wörter
20	2	843	191.171

Tab. 4-2: Gesamtanzahl der Trainings- und Testdaten für die Adaption

Die Parameter für die Adaption des Sprachmodells wurden im LM Workplace vorgegeben. Die Verfasserin hatte keinen Einfluss auf die Einstellungen. Für die beiden Fächer wurde das Training mit dem Koeffizienten 0,25 als Gewichtung für die Interpolation der Sprachmodelle durchgeführt.

Um die Adaption des Sprachmodells zu evaluieren, wurden Testdaten in LM Workplace geladen. Das Sprachmodell konnte entweder nur mit einem Textkorpus oder, in Kombination mit dem akustischen Modell, durch die Verwendung von Testdatensets, bestehend aus einer Audiodatei mit entsprechender Transkription, getestet werden. Zum Testen des Sprachmodells in Kombination mit dem akustischen Modell war es notwendig, die Audiospur vom Video zu trennen und zu extrahieren. Hierfür wurde der *Pazera-Audio-Extraktor*[62] verwendet. Die Audiodatei (wav-Format) und die dazugehörige Transkription (txt-Format), in einem Ordner komprimiert, konnten dann in LM Workplace zum Testen der Adaption hochgeladen werden. Für die Beurteilung der Adaption wurden die Werte von drei Evaluationsmaßen (WER, OOV-Rate und Perplexität) der Test-Sets im Basismodell mit denen des veränderten Modells verglichen.

[62] http://www.pazera-software.com/products/audio-extractor/

Experimente und Ergebnisse für das Fach Informatik

Das Training für das Fach Informatik erfolgte in zwei Phasen. In der ersten Phase wurde das Sprachmodell mit insgesamt 895.265, in der zweiten mit 461.273 Wörtern trainiert. In allen Dokumenten wurden die Themen *Datenanalyse, Datenzentren, Cloud Computing, Cloud Server, Server Chips und Softwareprogrammierung* behandelt. Die Abbildung 4-5 zeigt die Ergebnisse eines Tests nach der ersten Adaption mit der automatischen Transkription einer Podiumsdiskussion von 74 Minuten und 12.481 Wörtern. Der Abbildung kann entnommen werden, dass nur ein Teil der Evaluationsdaten automatisch erstellt wurde, da allein das Sprachmodell, ohne Kombination mit dem akustischen Modell, evaluiert wurde.

Base-Model-Evaluation (460856 Wörter im Vokabular)

Test-Set	Kommentar	#sent	#words	known W.	wcr	#sub	#del	#ins	wer	ser	#oov	%oov	PP	kn W PP
KIT_ComputerScience_EN		0	12481	11015							1466	11.75	1315.67	3370.95

Abb. 4-5: Evaluationswerte des Sprachmodells „Informatik“ nach der ersten Trainingsphase

Um auch die WER zu erhalten und diese in Relation mit den Perplexitätswerten stellen zu können, wurden in den nachfolgenden Experimenten das akustische und das Sprachmodell gemeinsam getestet. In der Tabelle 4-3 werden die Werte der Evaluationsmaßen OOV-Rate, Perplexität und WER für neue Test-Sets vor und nach der Adaption des Sprachmodells dargelegt.

Test-Set_ID	Known-Words Basis	Known-Words AM	OOV-Rate % Basis	OOV-Rate % AM	OOV-Count Basis	OOV-Count AM	PP Basis	PP AM	PP Variation	WER Basis	WER AM	WER Variation
Felten	12665	13312	0.45	0,44	57	59	1177,76	339	71%	88,6	78,3	10,3
KIT	11015	11307	11,75	11,48	1466	1466	1315,67	280	79%			
Stanford1a	12097	12095	0.6	0,65	73	79	805,925	205,99	74%	73,5	54,1	19,4
Stanford2a	10954	10954	1,19	1,19	132	132	992,217	254,441	74%	74	55	19
Stanford3	12491	12484	0.91	0,94	115	118	1536,36	293,286	81%	75,9	59,3	16,6
Stanford4	10873	10864	0.86	0.88	94	96	2233,05	367,151	84%	75,7	54,5	21,2
Stanford5	11511	11470	1,07	1,07	124	132	1863,62	355	81%	76,1	59	17,1
Stanford6	11730	11698	1,05	1,05	124	130	1408,09	301	79%	74,7	59,2	15,5
Stanford24	11517	11498	0.73	0,79	85	92	561	192,84	66%	73,5	53	20,5
Mittelwerte			**2,07**	**2,20**			**1321,54**	**287,65**	**76%**	**76,50**	**59,05**	**17,45**

Tab. 4-3: OOV-Rate, Perplexität und WER vor und nach der LM Adaption für „Informatik“

Die ersten zwei Spalten in der Tabelle 4-3 geben eine Übersicht über die Anzahl der im Vokabular des Spracherkenners vorhandenen Wörter, die auch im Test-Set gefunden wurden. Die Zahlen in der Spalte *KnownWords Basis* beziehen sich auf die Anzahl der im Vokabular gefundenen Wörter im Basissprachmodell. Diese werden der Anzahl von vorhandenen Wörtern in dem neugebildeten Sprachmodell in der Spalte *KnownWords AM* gegenübergestellt. Innerhalb von neuen Test-Sets korrespondieren 105.682 Wörter mit den im Vokabular des Spracherkenners vorhandenen Wortformen. Lediglich 2304 Wörter sind im Wörterbuch des Systems nicht bekannt. In den nächsten zwei Spalten werden die erzielten OOV-Raten für alle Test-Sets mit dem Basissprachmodell (Spalte *OOV-Rate Basis*) und mit dem adaptierten Sprachmodell (Spalte *OOV-Rate AM*) gegenüber gestellt. Obwohl der Anteil der OOV-Wörter relativ gering ist, ist die OOV-Rate im Durchschnitt nach der Adaption gestiegen. Der Grund dafür sind die hohen Werte des Test-Sets KIT, bei dem eine OOV-Rate von 11,48% (1466 Wörter) registriert wurde. Die kleine Verringerung der OOV-Rate nach der Adaption des Sprachmodells wird in der Abbildung 4-6 veranschaulicht[63].

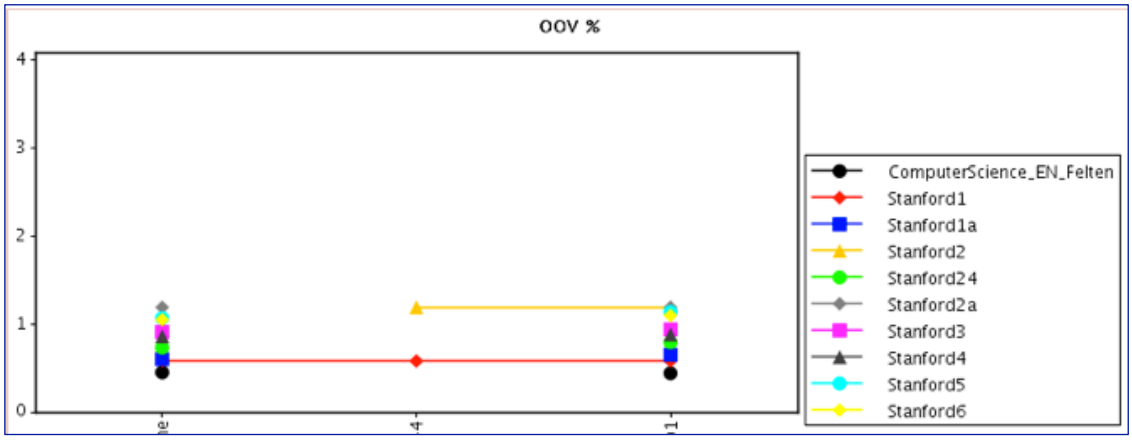


Abb. 4-6: OOV-Rate der Test-Sets „Informatik“ vor dem Training und nach Adaption

Nach der Adaption des Sprachmodells konnte eine Reduzierung der PP-Werte um 76% im Durchschnitt gemessen werden, wie die Werte in der Spalte *PP Variation* verdeutlichen. Die Perplexitätswerte für das Basis-Sprachmodell werden in der Spalte *PP Basis* den PP-Werten nach der Adaption gegenübergestellt (Spalte *PP AM)*. Durch Training konnten die PP-Werte von durchschnittlich 1321,54 auf

[63] Die in der Legende des Diagramms angezeigten Test-Sets Stanford 1 bzw. 2 waren fehlerhafte Dateien, die inhaltlich durch Stanford 1a bzw. 2a ersetzt wurden.

287,65 reduziert werden. Die Abbildung 4-7 veranschaulicht die Reduktion der Perplexität für die verwendeten Test-Sets nach der Adaption des Sprachmodells im Vergleich mit dem Basismodell.

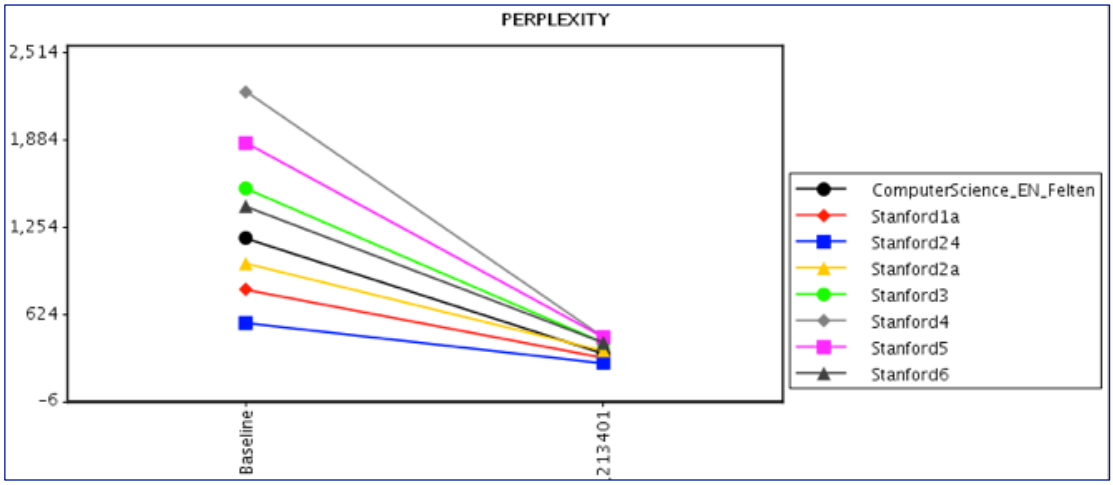


Abb. 4-7: Perplexitätswerte vor und nach der Adaption

In den Abbildungen 4-6 und 4-7 werden auf der Y-Achse die Werte der Evaluierungsmaße OOV-Rate bzw. Perplexität für die verwendeten Test-Sets mit dem Basissprachmodell (auf der X-Achse links) und mit dem adaptierten Sprachmodell (auf der X-Achse rechts) dargestellt.

Auch die Wortfehlerrate konnte durch das Training von durchschnittlich 76,50 (*WER Basis* für Test-Sets mit dem Basismodell) auf 59,05 (in der Spalte *WER AM* für das neue Sprachmodell) reduziert werden, es erfolgte somit eine Verringerung von 17,45. Das Diagramm in der Abbildung 4-8 veranschaulicht die erreichte Reduktion von WER für die verwendeten Testdaten, jeweils nach einer Trainings-phase, im Vergleich mit dem Basismodell.

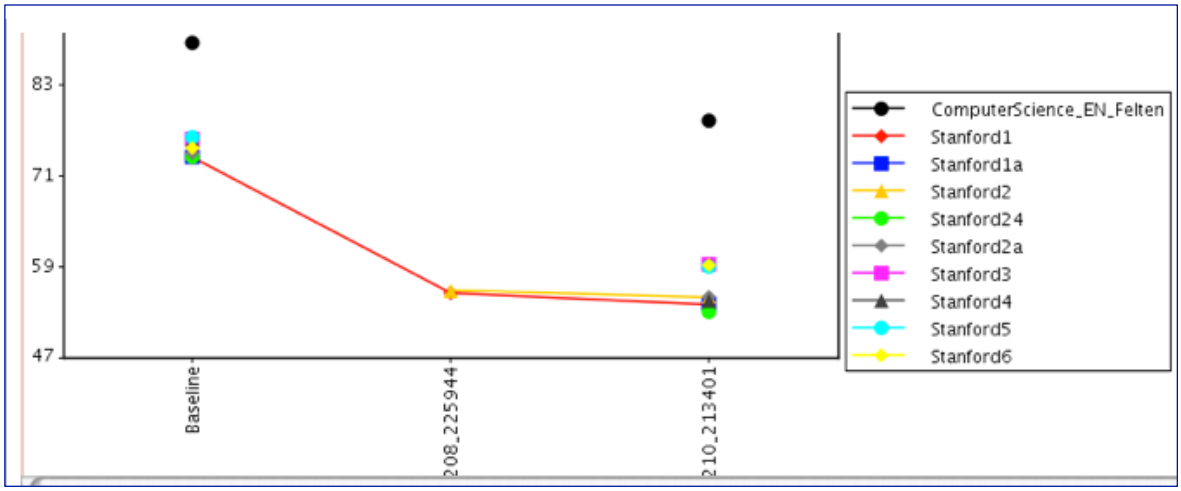


Abb. 4-8: Reduktion von WER nach beiden Trainingsphasen verglichen mit dem Basismodell

Der Tabelle 4-3 kann noch entnommen werden, dass bei den Test-Daten mit der Kennzeichnung Stanford 3 bis 5 die schlechtesten Werte für Perplexität und Wortfehlerrate registriert wurden (vgl. auch Abb. 4-7 bzw. 4-8), obwohl alle Stanford-Testdaten den gleichen Sprecher haben und ähnliche Themen behandelt werden. Neben den Werten für die Evaluation wurde auch die Anzahl der Wörter im Vokabular für das Basis-Sprachmodell von 222.660 Wörtern und für das durchgeführte Training von 461.273 Wörtern angezeigt. In der Anlage C7 werden die Evaluationsdaten und die Anzahl der Wörter pro Test-Set dargestellt.

Experimente und Ergebnisse für das Fach Technik

Das Training im Bereich Technik erfolgte ebenfalls in zwei Phasen. Zunächst wurde das Sprachmodell mit 928.956, dann mit 448.001 Wörtern trainiert. In allen Dokumenten wurden die Themengebiete Energietechnik und Robotik behandelt. Die Testdaten mit der Kennzeichnung *TED Energy* wurden für die Evaluation der ersten Adaptionsphase verwendet. Diese Test-Sets haben einen Umfang von insgesamt 17.691 Wörtern bei 110 Minuten. Nach der zweiten Trainingsphase wurde die Adaption erneut mit fünf Test-Sets (305 Minuten, 41.926 Wörter) getestet.

In der Tabelle 4-4 werden die Evaluationsergebnisse der Adaption des Sprachmodells nach der zweiten Trainingsphase für alle verwendeten Test-Sets dargestellt. Die ersten zwei Spalten in der Tabelle 4-4 geben eine Übersicht über die Anzahl der im Vokabular des Spracherkenners vorhandenen Wörter (*Known Words Basis*), die auch im jeweiligen Test-Set gefunden wurden. 62.949 Wörter in 12 Test-Sets korrespondieren damit. Lediglich 441 Wörter sind im Wörterbuch des Systems nicht bekannt. In den nächsten zwei Spalten werden die OOV-Raten für alle Test-Sets vor und nach der Adaption gegenüber gestellt. Die OOV-Rate konnte durch das Training nur geringfügig verbessert werden, von durchschnittlich 0,93 auf 0,76. Ein Grund dafür kann bei den Test-Sets mit der Kennzeichnung TED-Energy gesehen werden. Obwohl in diesen Videos Reden zum Thema Energietechnik vorkommen, dominiert deutlich die Allgemeinsprache.

Test-Set_ID	Known-Words Basis	Known-Words AM	OOV-Rate % Basis	OOV-Rate % AM	OOV-Count Basis	OOV-Count AM	PP Basis	PP AM	PP Variation	WER Basis	WER AM	WER Variation
Batteries Energy	7115	6986	0,7	1,19	50	84	329,159	258,657	21,42%	51,8	40,6	11,2
CellFree Energy	5592	5492	1,95	2,76	111	156	235,198	184,811	21,42%	42,7	37,7	5
Energy Conversion	7659	7700	1,01	0,44	78	34	194,561	174,471	10,33%			
FieldEnergy	7681	7711	0,9	0,41	70	32	317,148	259,266	18,25%	46,4	35,6	10,8
SpaceShuttle-Payloads	16407	16437	0,48	0,3	79	50	217,218	175,899	19,02%	39,1	36,1	3
TED_Energy	2001	2008	1,43	1,08	29	22	300,784	245,851	18,26%	32,1	23,7	8,4
TED_Energy2	594	594	1	1	6	6	289,510	228,465	21,09%	24,7	17,1	7,6
TED_Energy3	4666	4664	0,38	0,43	18	20	274,056	214,428	21,76%	48,6	37,7	10,9
TED_Energy4	964	969	1,03	0,51	10	5	300,300	277,442	7,61%	46,9	37,5	9,4
TED_Energy5	2899	2916	0,99	0,41	29	12	246,643	226,640	8,11%	48,4	39,4	9
TED_Energy6	4795	4814	0,64	0,25	31	12	202,976	163,654	19,37%	42,9	33,1	9,8
TED_Energy7	2648	2658	0,68	0,3	18	8	276,808	208,919	24,53%	45,1	34,6	10,5
Mittelwerte			**0,93**	**0,76**			**265,36**	**218,21**	**17,60%**	**42,61**	**33,92**	**8,69**

Tab. 4-4: OOV-Rate, Perplexity und WER vor und nach der Adaption für „Technik“

Dies wird durch die Anzahl der im Wörterbuch des Spracherkenners nicht bekannten Wörter bestätigt (Spalten *OOV-Count Basis* bzw. *AM*). Nur bei den ersten fünf Test-Sets in der Tabelle 4-4 (Spalte *OOV-Count AM*) ist die Anzahl der nicht vorhandenen Wortformen im Vokabular des Systems deutlich höher. Das Diagramm in der Abbildung 4-9 zeigt die geringe Reduktion der OOV-Rate. Hier

wird auch die deutliche Steigerung der OOV-Rate von 1,95 auf 7,76 nach der Adaption beim Test-Set *CellFreeEnergy* sichtbar.

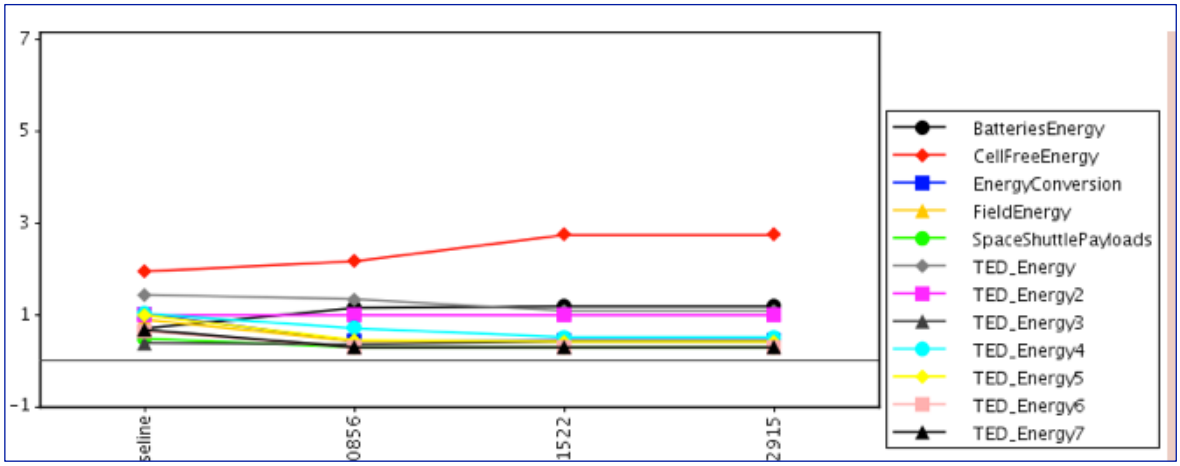


Abb. 4-9: Vergleich der OOV-Rates für 12 Test-Sets für „Technik“ vor und nach der Adaption

Nach der Adaption des Sprachmodells konnte eine Reduzierung der PP-Werte um 17,60% im Durchschnitt gemessen werden, wie die Werte in der Spalte *PP Variation* zeigen. Die Perplexitätswerte für das Basis-Sprachmodell werden in der Spalte *PP Basis* den Werten nach der Adaption gegenübergestellt (Spalte *PP AM)*. Durch Training konnten die PP-Werte von durchschnittlich 265 auf 218 reduziert werden. Die Abbildung 4-10 veranschaulicht die Reduktion der Perplexität für die verwendeten Test-Sets nach jeder Trainingsphase des Sprachmodells im Vergleich mit dem Basismodell.

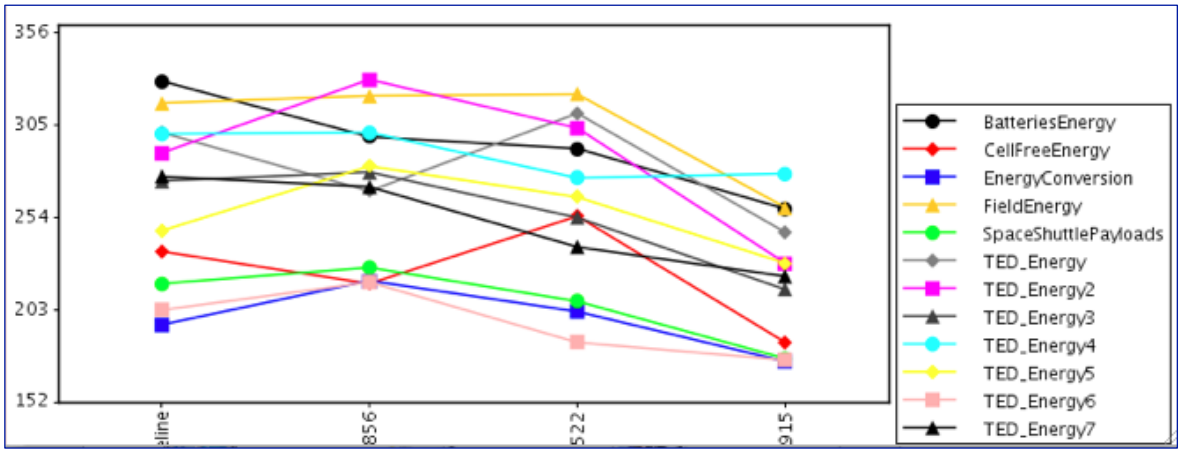


Abb. 4-10: Perplexitätsreduktion nach der Adaption des Sprachmodells für „Technik“

Auch hinsichtlich der Wortfehlerrate wurde durch das Training eine Reduktion der durchschnittlichen Werte von 42,61 (*WER Basis* für Test-Sets mit dem Basismodell) auf 33,92 (in der Spalte *WER AM* für das neue Sprachmodell) erreicht. Das Diagramm in der Abbildung 4-11 veranschaulicht die Entwicklung der Wortfehlerrate im Laufe des Trainings für alle verwendeten Testdaten jeweils im Vergleich mit dem Basismodell. Der Tabelle 4-4 kann entnommen werden, dass für den Test-Set *Energy Conversion* keine WER gebildet wurde. Die Daten fehlen deshalb auch im Diagramm. Der Gründe sind nicht bekannt.

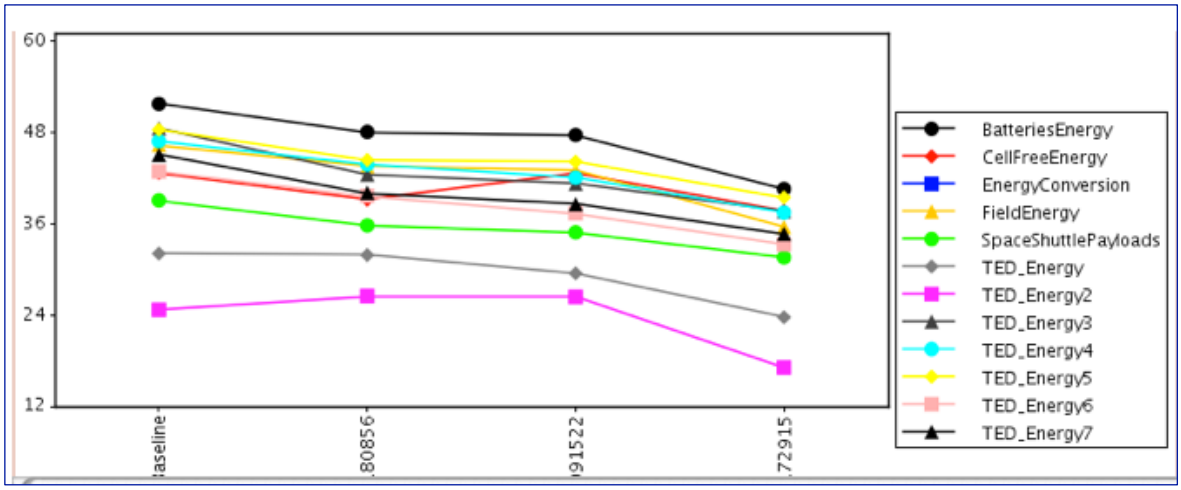


Abb. 4-11: Entwicklung der WER nach jeder Adaptionsphase im Vergleich mit dem Basismodell

4.4 Zusammenfassung und Schlussfolgerungen

In diesem Kapitel wurde gezeigt, dass das Web geeignete Ressourcen für das Training eines Sprachmodells bereithält, die zur Verbesserung automatischer Spracherkennung genutzt werden können. Menge, Art, Qualität und Relevanz der Quellen hängen erheblich von der Formulierung der Suchanfragen ab. Für das Training von Sprachmodellen und das Testen der Adaptionsergebnisse eignen sich besonders Transkriptionen gesprochener Sprache. Frei verfügbare Transkriptionen zu technischen Themen sind nur in geringen Mengen vorhanden. Die Ergebnisse der Experimente haben gezeigt, dass nicht die Menge der Daten für das Training eines Sprachmodells entscheidend ist, sondern die Qualität, vor allem die Qualität der Testdaten. Dies wird in der Tabelle 4-5 veranschaulicht. Obwohl für die Adaption von zwei Sprachmodellen (Informatik und Technik) etwa die gleiche

Datenmenge verwendet wurde, variieren die Werte der Evaluationsmaßen Wortfehlerrate, Perplexität und Out-of-Vocabulary-Rate stark.

Projekt	**Training**	**Ergebnisse Adaption**				
	Anzahl Wörter	OOV-Rate % AM	WER AM	WER Variation	PP AM	PP Variation
Informatik	1.356.538	2,2	59,05	17,45	287,65	76%
Technik	1.376.957	0,76	33,92	8,69	218,21	17,6

Tab. 4-5: Vergleich der Evaluationsmaße für Informatik und Technik

Der Prozess der Adaption eines Sprachmodells kann nur ein iterativer Prozess sein, während dessen die Ergebnisse laufend bewertet und davon abhängig die Trainingsstrategie geändert wird. Die Ergebnisse zeigen, dass nach jeder Trainingsphase sensible Verbesserungen eintreten. Jedoch ist die Überprüfung der Qualität von Test-Sets unerlässlich für die Beurteilung der Ergebnisse. Idealerweise erfolgt das Training nach der von Fügen (2009, S. 67–97) beschriebenen Methode, so dass nach durchgeführtem Training mit einem Datentyp die erzielten Adaptionsergebnisse gemessen werden können.

Die besten Ergebnisse der vorliegenden Untersuchung sind die Reduktion der Perplexität im Mittel um 76% und der Wortfehlerrate um 17,45 bei der Adaption des Sprachmodells „Informatik“. In der Tabelle 4-6 werden die Evaluationswerte für die Sprachmodelle „Informatik“ und „Technik“ gegenüber gestellt.

Test-Set_ID	Known-Words Basis	Known-Words AM	OOV-Rate % Basis	OOV-Rate % AM	OOV-Count Basis	OOV-Count AM	PP Basis	PP AM	PP Variation	WER Basis	WER AM	WER Variation
Informatik	104.853	105682	2,07	2,20	2270	2304	1321,54	287,65	76%	76,5	59,05	17,45
Technik	63.021	62.949	0,93	0,76	529	441	265,36	218,21	17,60%	42,61	33,92	8,69

Tab. 4-6: Evaluation der Adaption der Sprachmodelle „Informatik“ und „Technik“ im Vergleich

Die Ergebnisse der Experimente haben auch gezeigt, dass knapp 3% der Wörter aus den Test-Sets im Vokabular des Spracherkenners nicht vorhanden sind. Die unbekannten Wörter führen zu Fehlern bei der Spracherkennung und somit zu hohen Wortfehlerraten.

Im folgenden Kapitel wird untersucht, wie sich die Fachsprachen in Fachpublikationen und in Vorlesungsvideos unterscheiden und wie hoch der Anteil der Fachbegriffe in den beiden Medien ist.

5. Terminologie-Extraktion

Die Forschungsergebnisse eines Fachgebiets und die Erkenntnisse der Wissenschaftler werden der Fachwelt mittels Fachpublikationen mitgeteilt. Die unterschiedlichen Publikationstypen enthalten Daten, Fakten, Informationen und stellen somit Wissensquellen dar. Diese Informationsquellen liegen in schriftlicher Form vor und sind mittels natürlicher Sprache verfasst. Wissenschaftliche Texte werden in digitaler und strukturierter Form in vielen Datenbanken nachgewiesen. Solche digitalen, strukturierten Sammlungen stellen wichtige Quellen für die Identifizierung und Analyse von fachspezifischen Termini dar und für die Untersuchung der in der wissenschaftlichen Kommunikation verwendeten speziellen Sprachen.

Die Termini eines Wissensgebiets sind Bestandteile einer Fachsprache, die aufgrund der Eindeutigkeit der Termini einen präzisen Austausch über fachliche Konzepte ermöglicht. Die terminologischen Elemente sind Schlüsselbegriffe in Dokumenten. Die häufige Nutzung von Termini in wissenschaftlichen Texten bezeichnen Ahmad und Rogers (1997, S. 726) als „lexical density“, welche kennzeichnend für die Fachsprachen ist. Weitere Charakteristika der Fachsprachen sind, wie im Kapitel 2 gezeigt wurde, die spezielle Syntax, die häufige Nutzung von fachspezifischen Morphemen und die Verwendung bestimmter Muster bei der Bildung von Termini. Dazu gehört auch die signifikant hohe Verwendung von Nominalisierungen, Pluralformen sowie Mehrworttermini, mit denen die Entitäten in den Fachtexten referenziert werden.

Diese Mehrworttermini stellen Phrasen dar, welche meistens aus Substantiven und Adjektiven, selten auch mit Präpositionen, nach unterschiedlichen Mustern gebildet werden. Diese Charakteristika der Terminologie und Fachsprachen sind die Grundlagen für die Identifizierung und Extraktion von fachspezifischen Termini durch die Anwendung von statistischen, linguistischen oder hybriden Methoden.

In dieser Arbeit wird die Terminologie-Extraktion dazu genutzt, um die Häufigkeit der Fachtermini in wissenschaftlicher Fachliteratur und Vorlesungsvideos zu ermitteln und zu vergleichen.

5.1 Ausgewählte Methoden zur Terminologie-Extraktion

Die folgenden Methoden zur Terminologie-Extraktion bieten einen Überblick zum Forschungsstand und lassen sich in drei Hauptrichtungen einteilen: statistische, linguistische und hybride Verfahren. Bei den hybriden Verfahren handelt es sich um eine Kombination aus Sprachstatistik und linguistischen Analysen.

5.1.1 Statistische Methoden

Durch den Einsatz von statistischen Analysen können in großen Dokumentensammlungen relevante Merkmale und Strukturen effizient identifiziert und selegiert werden. Hierfür werden die Häufigkeitsanalyse von Wortformen und die Differenzanalyse verwendet. Andere Methoden, wie die Analyse von signifikanten Kollokationen und Kookkurrenzen, betrachten die statistischen Abhängigkeiten zwischen den Wortformen unter Berücksichtigung von sog. syntagmatischen und paradigmatischen Relationen, welche zwischen den Wortformen bestehen (vgl. Heyer et al., 2006, S. 85–87; Granitzer, 2006, S. 444–445). Abschließend wird eine der meist benutzten Methoden vorgestellt, die aus dem Bereich Information Retrieval stammt und die Kombination zweier Messgrößen darstellt: TF (Term Frequency) und IDF (Inverse Document Frequency).

Die Häufigkeitsanalyse und das Zipfsche Gesetz

Fachtermini können mittels statistischer Verfahren in Texten identifiziert werden. Durch die Anwendung des Zipfschen Gesetzes, das den Zusammenhang zwischen Rang und Häufigkeit aufzeigt, können unterschiedliche Berechnungen mit einem und Aussagen über einen Text gemacht werden (vgl. Heyer et al., 2006, S. 87–93). Geeignete Beispiele sind:

- Der Umfang eines benötigten Textes, in dem eine vorgegebene bestimmte Anzahl von Wortformen genau n-mal vorkommt, kann geschätzt werden.

- Die relative Häufigkeit von Wortformen, die n-mal im Text vorkommen, kann unter Heranziehung der Gesamtanzahl von Wortformen im Text berechnet werden.
- Die Relation zwischen Textgröße und Umfang des Vokabulars kann geschätzt werden.

Bei einem Text wird der Zusammenhang zwischen Rang und Häufigkeit erkennbar, wenn man die Wortformen des Textes absteigend nach ihrer absoluten Häufigkeit *n* ordnet und den Platz einer Wortform in der Liste mit Rang *r* bezeichnet. Wird der Wert der Häufigkeit *n* mit dem Wert des Rangs *r* multipliziert, ergibt sich jeweils ein konstantes Ergebnis: $r \cdot n \approx k$ (vgl. Heyer et al., 2006, S. 87–88). Daraus ergibt sich folgende Feststellung: Je höher die Häufigkeit ist, desto niedriger der Rang einer Wortform in der Liste der Wortformen. Die Häufigkeitsverteilung zeigt, dass wenige Wortformen sehr häufig vorkommen und viele Wörter weniger oft vorkommen.

Die Ergebnisse einer älteren Untersuchung von Hoffmann (1988, S. 62 zit. in Witschel, 2004, S. 17) konkretisieren die oben genannte Feststellung:

- Die Funktionswörter (Artikel, Präpositionen, Konjunktionen), welche eine syntaktische Funktion im Text haben, sind unter den häufigsten Wörtern sowohl in Fachsprachen als auch in der Allgemeinsprache vertreten,
- bestimmte Nomina, die in einer Fachsprache häufig vorkommen, treten in der Allgemeinsprache wesentlich seltener auf oder kommen gar nicht vor,
- häufig auftretende Termini sind charakteristisch für spezielle Fachgebiete.

Ein Beispiel mit den zehn häufigsten Wörter aus dem Projekt „Deutscher Wortschatz“ findet sich bei Heyer et al. (2006, S. 88) und zeigt, dass die häufigsten Wortformen die kurzen Funktionswörter sind. Diese Ergebnisse verdeutlichen das Verhältnis zwischen Allgemeinsprache und Fachsprachen, welches in der Abbildung 5-1 grafisch veranschaulicht wird.

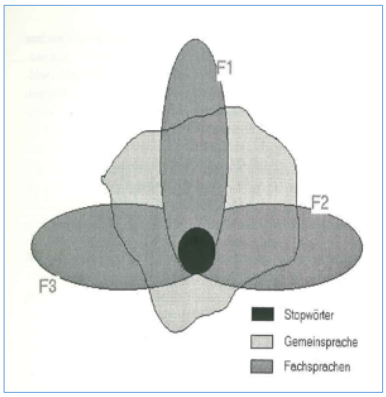


Abb. 5-1: Verteilung lexikalischer Elemente in Gemeinsprache und Fachsprachen.
Quelle: Witschel, 2004, S. 69

Die absolute Häufigkeitsverteilung der Wortformen im Text bzw. in einem Textkorpus (Sammlung von Texten) wird benötigt, um durch Vergleich der unterschiedlichen Verteilungen der Wortformen innerhalb von zwei Korpora die diskriminierenden Termini zu ermitteln und zu extrahieren. Es handelt sich dabei um eine statistische Methode, die Differenzanalyse genannt wird und auch in dieser Arbeit bei der Terminologie-Extraktion angewendet wird. Die Vorgehensweise bei der Differenzanalyse wird im folgenden Abschnitt beschrieben.

Die Differenzanalyse

Die Anwendung der Differenzanalyse erfordert die Bereitstellung von zwei Korpora: einen Analyse- und einen Referenzkorpus. Aus den Texten des Analysekorpus werden die diskriminierenden Termini extrahiert. Der Analysekorpus muss somit aus fachspezifischen Texten gebildet werden. Der Referenzkorpus besteht im Gegensatz dazu aus Texten, in denen die Allgemeinsprache verwendet wird. Im nächsten Schritt wird die Häufigkeitsanalyse für einzelne Wortformen oder für Wortformkombinationen der beiden Korpora durchgeführt. Die berechneten Werte für die beiden Korpora werden miteinander verglichen, indem die Mengendifferenz der fachspezifischen und der allgemeinsprachlichen Wortformen gebildet wird (vgl. Heyer et al., 2006, S. 95-98).
Auf dieser Grundlage kann eine Einteilung der Wortformen in vier Klassen erfolgen. Klasse 1 enthält die Wortformen, die im Referenzkorpus nicht vorkommen. Bei diesen Wortformen handelt es sich um die zu ermittelnden fachspezifischen Termini aus dem Analysekorpus. Der Klasse 2 werden alle Wortformen

zugeordnet, welche im Analysekorpus häufiger vorkommen als im Referenzkorpus. Die Klasse 3 enthält alle Wortformen, welche in beiden Korpora etwa gleiche Häufigkeitswerte haben. Bei diesen Wortformen handelt es sich um Funktionswörter bzw. um allgemeine Begriffe. Die letzte und vierte Klasse enthält alle Wortformen, welche im Fachkorpus seltener als im Referenzkorpus vorkommen. Diese Wortformen transportieren aber keinen Informationsgehalt und gelten daher nicht als Fachterminologie. Diese Klasseneinteilung stellt die Basis für die weitere Analyse dar, bei der die Wortformen der Klasse 1 und 2 betrachtet werden sollen.

Bei den Wortformen aus der Klasse 2 handelt es sich wahrscheinlich um Fachtermini. Heyer et al. (2006, S. 95–98) empfehlen ein Maß einzusetzen, welches die Identifizierung der tatsächlichen Fachtermini ermöglicht: Eine Mindestdifferenz der Häufigkeitsklassen, die als Schwellenwert definiert wird. Dafür berechnen Heyer et al. (2006, S. 97) zunächst die relative Häufigkeit der Wortformen aus dem Analysekorpus. Dann werden insbesondere diejenigen Wortformen betrachtet, deren Häufigkeit im Analysekorpus im Vergleich mit der relativen Häufigkeit im Referenzkorpus wesentlich höher ist.

Eine Mindestfrequenz zusammen mit einer zusätzlichen statistischen Prüfgröße *P* verwendet auch Witschel (2004, S. 17–19; S. 57) bei der Implementierung eines von ihm entwickelten Extraktionsverfahrens. Diese Prüfgröße *P* ähnelt dem oben beschriebenen Schwellenwert von Heyer et al. (2006, S. 97). Witschel bezweckt damit, diejenigen Wortformen herauszufiltern, deren Frequenzen signifikant höher liegen als der Wert der relativen Häufigkeit der jeweiligen Wortformen im Referenzkorpus. Auch Witschel bedient sich somit der relativen Häufigkeit, um die Wahrscheinlichkeit des Auftretens einer Wortform in einem Fachtext zu schätzen und geht dabei von der Nullhypothese aus. Nach der Nullhypothese ist die Wahrscheinlichkeit, dass ein Wort *w* im Fachtext vorkommt, gleich der aus dem Referenzkorpus geschätzten Wahrscheinlichkeit *p* (vgl. Witschel, 2004, S. 57).

Witschel verwendet unterschiedliche statistische Tests (Likelihood-ratio-Test, Poisson-Verteilung, Häufigkeitsquotient) und vergleicht, welcher Test die besten Ergebnisse hinsichtlich Precision und Recall liefert (Witschel, 2004, S. 58–61; S.

73–77). Fachspezifische Einworttermini filtert Witschel mittels Differenzanalyse, indem er alle Wortformen auswählt, „deren Frequenz oberhalb eines gewissen Mindestwertes liegt und für die der Wert einer der oben beschriebenen Prüfgrößen einen Schwellenwert überschreitet“ (Witschel, 2004, S. 62).

Die Analyse von Kollokationen und signifikanten Kookkurrenzen

In der Literatur finden sich in der Vorgehensweise unterschiedliche Hinweise auf die Extraktion von Kollokationen mittels statistischer Verfahren, z.B. bei Manning und Schütze (2001, S. 151–189), Heid (1997, S. 788–808), Frantzi und Ananiadou (1996, S. 41–46, zit. in Pazienza, 1998, 187) und bei Granitzer (2006, S. 445).
Bezugnehmend auf den linguistischen Strukturalismus, der von dem Schweizer Linguisten Ferdinand de Saussure Ende des 19. Jahrhunderts geprägt wurde, definieren[64] Heyer et al. (2006, S. 20–25) die „signifikanten Kookkurrenzen“ und Witschel (2004, S. 24–25) die „Nachbarschaftskollokationen“ gleichermaßen: Zwei Wortformen stehen in syntagmatischer Relation zueinander, wenn sie in einem lokalen Kontext gemeinsam auftreten. Dabei ist es von Bedeutung, dass die beiden Kookkurrenten nicht gelegentlich zusammen auftreten, sondern statistisch signifikant oft. Es wird unterschieden zwischen Nachbarschaftskookkurrenzen (wenn die beiden Wortformen unmittelbar benachbart auftreten) und Satzkookku-r-renzen (wenn die Wortformen innerhalb eines Satzes vorkommen), wobei die ersteren auch Satzkookkurrenzen sind, da zwei benachbarte Wortformen im gleichen Satz stehen (vgl. Heyer et al., 2006, S. 136–137).

Für die Ermittlung von signifikanten Kookkurrenzen wird üblicherweise die Häufigkeitsanalyse verwendet, bei der das zuvor extrahierte Merkmal, z.B. die Wortgruppe der Länge *n*, auf das Vorkommen der unterschiedlichen Wortformen untersucht wird. Ein Beispiel hierfür ist das Verfahren von Frantzi und Ananiadou (1996, S. 41–46, zit. in Pazienza, 1998, 187), die in ihrer Arbeit die Extraktion von Kollokationen mittels statistischer Analysen fokussiert haben. Deren statistische Analyse nutzte die Länge von n-grams, ihre absolute Häufigkeit im Korpus, die

[64] Vgl. dazu die Definitionen des Begriffs *Kollokation* von Wright (1997, S. 15–16), Heid (1997, S. 788–789) und Witschel (2004, S. 24–25) im Kapitel 2.1.2, Abschnitt *Komposition und Phrasenbildung.*

Häufigkeit der längsten Wortform sowie Kollokationen und ihre Anzahl. Bei solchen Verfahren ist eine Vor- und Nachbearbeitung notwendig, um nur semantisch sinnvolle Wortkombinationen zu ermitteln (vgl. Granitzer, 2006, S. 445; Heid, 1997, S. 804–805).

Daran knüpfen auch Manning und Schütze (2001, S. 153–157) an, indem sie empfehlen, die einfache die Methode, Kollokationen in einem Korpus mittels Berechnung ihrer Häufigkeit zu identifizieren, durch einen Part-of-Speech-Filter, so wie er von Justeson und Katz (1995, S. 9–27)[65] entwickelt wurde, zu verbessern und begründen dies mit den erzielten, korrekten Experimentergebnissen. Die Häufigkeitsanalyse ergibt gute Resultate bei festen Kollokationen. Variiert der Abstand zwischen Wortformen, die in einem Text häufig zusammen vorkommen, muss die Verteilung dieser Abstände zwischen zwei Wörtern im Korpus berechnet werden und nach den Paaren mit den niedrigsten Abweichungen in der Länge der Wortabstände gesucht werden.

Das Verhältnis zwischen den relativen Häufigkeiten (*relative frequency ratios*) ist nach Manning und Schütze (2001, S. 175–176) das beste Verfahren für die Ermittlung und den Vergleich von signifikanten Kollokationen innerhalb von zwei oder mehreren Korpora. Der Vorteil liegt im Auffinden themenspezifischer Kollokationen. Dabei werden für Kollokationen aus zwei Wörtern die relativen Häufigkeiten in zwei unterschiedlichen Korpora berechnet und die ermittelten Frequenzwerte jeder Kollokation durch Division in Beziehung gesetzt. Das Ergebnis der Division ist die Grundlage für das Ranking der Kollokationen. Dieses Verfahren wird auch in der vorliegenden Arbeit bei der Analyse der extrahierten Wortformen verwendet (vgl. dazu Kap. 5.3.4).

Bei der Ermittlung von Wortformen, die häufig zusammen auftreten, können auch die paradigmatischen Relationen zwischen zwei oder mehreren Wortformen analysiert werden. Die paradigmatische Relation zwischen zwei oder mehreren Wortformen ist gegeben, wenn die globalen Kontexte der Wortformen hinsichtlich der inhaltlichen Zusammenhänge, der Semantik und der Syntaktik ähnlich sind.

[65] Das von Justeson und Katz entwickelte Verfahren wird im Kapitel 5.1.3 beschrieben.

Die globalen Kontexte der Wortformen müssen auf ihre Ähnlichkeit überprüft werden, was die Heranziehung eines Ähnlichkeitsmaßes erfordert, z.B. Tanimoto-Ähnlichkeit oder Cosinus-Maß (vgl. Heyer et al., 2006, S. 25–30).

Wartena, Brussee und Slakhorst (2010, S. 54–58) nutzen die Informationen über die paradigmatischen Relationen von Wortformen aus der Analyse der Verteilung von Kookkurrenzen zur Extraktion von relevanten Schlüsselwörtern aus Texten. Die Autoren gehen von der Hypothese aus, dass semantisch ähnliche Wörter in ähnlichen Kontexten gemeinsam mit den gleichen Wörtern auftreten. Bei der entwickelten Methode wird ein Vergleich der Kookkurrenzen der jeweiligen Termini mit der Wortformenhäufigkeit in allen Dokumenten eines Korpus vorgenommen. Dafür wurde die Kookkurrenz-Verteilung eines Wortes als gewichteter Durchschnitt der Wortverteilungen in allen Dokumenten definiert, in denen das Wort vorkommt. Die semantische Ähnlichkeit von zwei Termini wird gemessen, in dem die Verteilung ihrer Kookkurrenzen berechnet und auf Ähnlichkeiten geprüft wird. Precision und Recall-Werte von statistischen Verfahren können durch die Berücksichtigung von semantischen Relationen zwischen den Wortformen erhöht werden.
Die Kookkurrenz-Analyse wird in erster Linie beim Information Retrieval angewendet. Bei Heyer et al. (2006, S. 141–149) finden sich einige Anwendungsbeispiele für die automatische Extraktion von Daten aus Texten sowie für die Extraktion von semantischen Relationen, die zwischen signifikanten Kookkurrenzen bestehen.

Term frequency (TF) und Inverse document frequency (IDF)

Bei diesem aus dem Information Retrieval stammenden Verfahren wird eine Gewichtung eines Merkmals vorgenommen. Jeder Wortform in einem Dokument wird zunächst eine Gewichtung zugewiesen, die ihrer Frequenz im Dokument entspricht. Das wird folgendermaßen repräsentiert: $tf_{t,d.}$ Die Idee von Inverse document frequency ist es, die hohe Gewichtung zu reduzieren, welche manche Wortformen aufgrund der Häufigkeit ihres Vorkommens erhalten, indem die Anzahl der Dokumente berechnet wird, die in einer Sammlung vorhanden sind und

die jeweilige Wortform enthalten. Die Inverse document frequency eines Wortes wird wie folgt definiert:

$$idf_{\mathrm{t}} = log\,\frac{N}{df\ t}$$

wobei *N* die Gesamtanzahl der Dokumente in einer Sammlung ist und df_t die Dokumenthäufigkeit, welche die Anzahl der Dokumente wiedergibt, die die jeweilige Wortform enthalten (vgl. Manning, Raghavan & Schütze, 2008, S. 107–109). Das ergibt bei einer selten vorkommenden Wortform einen hohen und bei häufig benutzten Wortformen einen eher niedrigen idf-Wert. Das TF-IDF-Gewichtungsschema wird durch die Multiplikation der beiden o.g. Werte erzeugt: $tf_{t,d}$ x $idf_{t.}$ Es ermöglicht die Zuweisung einer dreistufigen Gewichtung der Wortformen:

- Die höchste Gewichtung erhalten die Wortformen, die sehr häufig, jedoch nur in einer kleinen Anzahl von Dokumenten vorkommen. Dies lässt sich wie folgt erklären: die hohe Frequenz einiger Wortformen zeigt deren Relevanz für die Beschreibung des Dokumentes und die Tatsache, dass diese Wortformen nur in wenigen Dokumenten vorkommen, hebt diese Dokumente innerhalb der Dokumentensammlung hervor (vgl. Granitzer, 2006, S. 446; Witschel, 2004, S. 37).
- Eine niedrigere Gewichtung erhalten die Wortformen, die relativ häufig in einem oder in vielen Dokumenten vorkommen, da dies ein Zeichen für niedrigere Relevanz als diskriminierender Terminus ist.
- Die niedrigste Gewichtung erhalten die Wortformen, die in allen Dokumenten vorkommen, das sind Wortformen aus der Allgemeinsprache.

Nach Witschel (2004, S. 37) entspricht die TF-IDF-Methode grob dem Prinzip der Differenzanalyse, wenn sie in Verbindung mit einer festgelegten Mindestfrequenz und einer statistischen Prüfgröße durchgeführt wird. Allerdings lässt sich das TF-IDF-Gewichtungsschema nur bei heterogenen Dokumentensammlungen anwenden und ist für Terminologie-Extraktion mit einem Referenzkorpus nicht geeignet. Das TF-IDF-Gewichtungsschema eignet sich für die Identifizierung von spezifischen Termini, die für die Beschreibung der Dokumente einer Sammlung relevant sind. Es hilft bei der Beurteilung von *term specificity* oder *termhood* eines Begriffes (vgl. Pazienza, 1998, S. 193). Die Fachspezifik eines Terminus (*termhood*) und ihre Bedeutung bei der Auswahl von Termini wird im folgenden Kapitel im

Abschnitt *Suche nach morphologischen Mustern* thematisiert. Es kann bei der Zuweisung von geeigneten Schlagwörtern oder beim Aufbau von Indices verwendet werden (vgl. Wartena, Brussee & Slakhorst, 2010, S. 54; Pazienza, 1998, S. 193).

5.1.2 Linguistische Methoden

Die linguistischen Verfahren bewirken die Extraktion von Mehrworttermini, in dem nach syntaktischen und morphologischen Mustern, mittels regulärer Ausdrücke oder durch den Einsatz von *Chunkern*, gesucht wird. In der Regel werden die zu analysierenden Dokumente zuerst mit einem Part-of-Speech-Tagger nach Wortarten annotiert.

Die regulären Ausdrücke stellen eine „Kunstsprache“ dar (vgl. Heyer et al., 2006, S. 227), die es ermöglicht, Muster zu formulieren, nach denen gesucht werden soll. Für die Formulierung der regulären Ausdrücke können die Atome, aus denen diese Sprache besteht, mit Hilfe von sieben unterschiedlichen Operatoren verknüpft werden. Atome können einzelne Zeichen, Zeichenfolgen oder Klassen sowie Bereiche von Zeichen sein (z.B. (a-z), [aeiouäöü] oder [ing]). Durch den Einsatz von Wildcards, die den Atomen mit Ausnahme des Escape-Zeichens „\“ nachgestellt werden, können beliebig oft vorkommende Zeichenfolgen gesucht werden.

Suche nach syntaktischen Mustern

Mit diesem Verfahren erfolgt die Suche nach speziellen Mustern, welche die Struktur des Satzes berücksichtigen. Wie oben erwähnt, werden die Wortartinformationen mit Hilfe eines Part-of-Speech-Taggers erzeugt. Part of Speech (POS) Tagging bedeutet, dass jedes einzelne Satzelement mit entsprechenden Informationen zu seiner Art versehen wird. Diese Tags sind bei der Suche nach Wortarten und benannten Entitäten in Texten nützlich. Die Aufgabe des Taggers ist, jedem Satzelement eine Annotation mit der entsprechenden Wortart zuzuweisen. Es gibt acht standardisierte POS-Tags: Adjektive, Adverbe, Konjunktionen, Artikel, Nomen, Pronomen, Präpositionen und Verben. Zusätzlich gibt es zwei weitere

POS-Tags für Satzzeichen und Interjektionen (vgl. Konchady, 2008, S. 192–194). Mit Hilfe der regulären Ausdrücke kann nach Wortformen oder Teilen davon, nach einzelnen POS-Tags sowie nach Paaren aus Wortform und dem dazugehörigen POS-Tag gesucht werden (vgl. Heyer et al., 2006, S. 233–238).

Im folgenden Abschnitt wird der Ansatz von Arppe dargestellt, der einen Chunker für Nominalphrasen verwendet, um aus mit POS-Tags annotierten Dokumenten Termini zu extrahieren.

Arppe (1995) entwickelte das Programm NPtool, das die Identifikation und Extraktion von englischen Nominalphrasen ermöglicht. Dabei wurden alle korrekten Teile der ermittelten Maximal-Phrasen aufgelistet, geordnet nach dem grammatikalischen Kopf der Phrase und deren Häufigkeit. Um echte Termini zu extrahieren, wurden die POS-Muster der potentiellen Termini untersucht. Er zeigte, dass mit acht POS-Mustern ca. 90% der Termini gefunden werden können. Die wichtigsten POS-Muster sind nach Arppe: AN, NN, N, NPN, ANN, NPAN, NPNN, ANPN, geordnet nach dem Recall des jeweiligen Musters, wobei N für Nomen, A für Adjektiv und P für Präposition stehen. Arppe stellte fest, dass die Precision für alle Muster zwischen 15–40% variiert, d.h. in diesem prozentuellen Anteil der extrahierten POS-Muster-Gruppen befinden sich echte Termini.

Eine der Schlussfolgerungen war, dass die kurzen Termini, einfache Nomen und aus zwei Wörtern gebildete Komposita (entsprechend den Mustern AN und N) ca. 80% des Vorkommens in Texten ausmachen. Nominalphrasen aus mindestens zwei Elementen haben eine Wahrscheinlichkeit von ca. 70%, echte Termini zu sein. Die Ergebnisse dieser Untersuchung geben Hinweise, mit welchen POS-Mustern gute Ergebnisse bei der Extraktion von Termini erzielt werden können. Der Vorteil der syntaktischen Analyse von Nominalphrasen (auch Chunking genannt) liegt, verglichen mit der einfachen Extraktion von POS-Mustern, darin, dass nur zusammenhängende Phrasen mit den gesuchten Mustern extrahiert werden (vgl. Witschel, 2004, S. 43).

Suche nach morphologischen Mustern

Die Morphologie untersucht die Art und Weise, wie Morpheme zu Wortformen kombiniert werden. Ein Morphem ist das kleinste bedeutungstragende Element der Sprache. Es wird zwischen Basismorphemen, Flexiven und Derivativen unterschieden. Charakteristisch für die Fachsprachen ist, neben dem fachspezifischen Vokabular, eine besondere Syntax, die durch die Nutzung von speziellen Wortbildungsmustern entsteht. Einen Überblick hierzu geben die Darstellungen im Kapitel 2.1.2. Die Identifizierung von Morphem-Mustern ist für die Extraktion von Fachtermini und bei der Erkennung semantischer Zusammenhänge nützlich. Die Kenntnis der typischen Morpheme des jeweiligen Fachgebietes ist somit unerlässlich.

Morpheme innerhalb von Wortformen können mittels regulärer Ausdrücke gefunden werden. Fachspezifische Morpheme können auch mittels Differenzanalyse gefiltert werden (vgl. Kap. 5.1.1). In der Regel wird eine Grundformreduktion der Wortformen vorgenommen. Dabei werden die Flexive von den Wortstämmen abgeschnitten. Nach Heyer et al. (2006, S. 242) ist die Aufgabe der Grundformreduktion je nach Sprache unterschiedlich schwer. Die o.g. Autoren geben dort das Beispiel der englischen Sprache an, bei der aufgrund der wenigen Flexive die Grundformreduktion einfach zu realisieren ist. Zur Identifizierung von fachspezifischen Morphemen wird oft auch eine Kompositazerlegung vorgenommen, um domänenspezifische Basismorpheme zu finden, weil Komposita fachspezifische Termini sein könnten. Basismorpheme helfen bei der schnellen Auffindung der Termini (vgl. Heyer et al., 2006, S. 238–245; Witschel, 2004, S. 23).

Basierend auf den existierenden linguistischen Verfahren bezeichnen Pazienza, Pennacchiotti und Zanzotto (2005, S. 258) ein Verfahren zur Extraktion von Termini als geeignet, das aus folgenden Phasen besteht: Analyse eines Korpus und Zuweisung von POS-Tags, Identifizierung und Extraktion von potentiellen Termini mittels festgelegter Regeln, Zusammenführen aller Variationen eines Terminus aufgrund ihrer semantischer Bedeutung sowie das Nutzen eines zusätzlichen linguistischen Verfahrens zum Filtern der richtigen Termini. An dieser Stelle weisen die drei o.g. Autoren auf die Notwendigkeit hin, die Definition von

„termhood“ (Fachspezifik der Termini) so in den Prozess einzubinden, dass die Auswahl von richtigen Termini anhand ihrer syntaktischen Struktur möglich wird.

Die Probleme der formalen Variation, die bei genormten Terminologien nicht auftritt, und der Trennbarkeit mancher Wortgruppen (vgl. Witschel, 2004, S. 26; Justeson & Katz, 1995, S. 10), sind Gründe für den Einsatz anderer Verfahren in Kombination mit den linguistischen Methoden. Ein Verfahren, das linguistische und statistische Methoden sowie Wissensquellen aus verschiedenen Bereichen mit einander kombiniert, wird hybrid genannt (vgl. Heyer et al., 2006, S. 247). Einige Möglichkeiten der hybriden Verfahren werden im folgenden Kapitel dargestellt.

5.1.3 Hybride Methoden

Bei der Anwendung eines hybriden Verfahrens für die Terminologie-Extraktion wird zunächst der Korpus mittels einer linguistischen Methode analysiert und anschließend werden statistische Verfahren auf die Menge der potentiellen Termini angewandt, die das Ergebnis der ersten Phase sind. Folgende Verfahrenskombinationen sind möglich (vgl. Pazienza et al., 2005, S. 263–264):

- Zuerst werden Nominalphrasen extrahiert und anschließend mit Hilfe der Häufigkeitsberechnung aus deren Elementen die richtigen Termini selektiert.
- Aus einem Korpus werden potentielle Termini mittels syntaktischer POS-Muster identifiziert und selektiert, die danach mit Hilfe von unterschiedlichen statistischen Berechnungen (Log-Likelihood Ratio, Mutual Information und Häufigkeitsanalyse) gefiltert werden.
- Alternativ können die potentiellen Termini in einem Korpus mit Hilfe von regulären Ausdrücken ermittelt werden, um die ausgewählten Wortformen danach einem Rankingverfahren aufgrund der Häufigkeitsanalyse zu unterziehen.

Die Anwendung von linguistischen Verfahren als erste Phase eines hybriden Modells ist nicht zwingend. Es ist möglich, zuerst „einfache“ Termini aufgrund ihrer Frequenz auszuwählen, um auf dieser Grundlage komplexe Termini mittels linguistischer und statistischer Filter zu ermitteln.

Die linguistische Analyse eines Korpus kann durch semantische Informationen aus Thesauri vertieft werden. So können solche Informationen zusammen mit den Ergebnissen von statistischen Messungen genutzt werden, um die potentiellen Termini zu ermitteln. Als statistisches Maß wird eine Kombination aus dem C-Maß (*C-value*) und dem sog. „Kontext-Faktor“ empfohlen. Der Kontext-Faktor berücksichtigt die semantischen, syntaktischen und statistischen Merkmale des Kontextes, in dem die potentiellen Termini vorkommen (Pazienza et al., 2005, S. 264). Das C-Maß ist eine linguistisch basierte Messgröße zur Ermittlung der „termhood“ für Mehrworttermini. Bei der Berechnung des C-Maßes werden unterschiedliche statistische Informationen über die Termini verwendet. Die Definitionen von „termhood“ und C-Maß, nebst Berechnungsformel, sind bei Pazienza et al. (2005, S. 259 u. 262–263) zu finden.

Zwei Ansätze für hybride Verfahren werden hier vorgestellt. Der erste Ansatz zur automatischen Extraktion von multilingualen Termini stammt von Bernhard (2006, S. 171–174) und basiert auf der Kombination eines linguistischen Verfahrens zur Identifikation von klassischen Morphemen mit zwei statistischen Methoden (Häufigkeitsanalyse und Log-Likelihood Ratio) sowie einem Clustering-Verfahren. Dort werden englische und französische Termini aus Korpora extrahiert, die mit Dokumenten aus den Bereichen der Medizin und Vulkanologie gebildet wurden. Im Fachgebiet Medizin werden Termini mittels Derivativen sowie altsprachlichen Wortstämmen gebildet. Dabei gibt es einige typische Morpheme, die sehr häufig auftreten. Als klassische Präfixe nennt Bernhard *extra-* sowie *anti-* und als Suffixe *-ism.* Weitere kombinatorische Elemente werden von Bernhard (2006, S. 171) „initial combining forms“ (*hydro-, pharmaco-*) sowie „final combining forms“ (*-graphy, -logy*) bezeichnet. Sager et al. (1980, S. 260) beschreiben die Art dieser kombinatorischen Elemente durch die Erklärung, warum altsprachliche Wortstämme durch die häufige Nutzung einen Präfixcharakter entwickelt haben:

> In special languages certain neoclassical stems are so frequently used that they can be considered prefixes from the functional and structural point of view, [...]. Stems like *carb-, cycl-, electr-, ferr-* and *hom-* develop prefix character by adding *o* to the stem, though the stem alone also occurs in middle or final position of compounds, e.g. [...] *hydrosphere, hydrogen, anhydride, hydroelectric* [...]

Zunächst wird eine Häufigkeitsliste mit allen aus dem Korpus extrahierten Wortformen erstellt. Durch den Einsatz des regulären Ausdrucks werden Präfixe gefunden, die eine Länge von mindestens vier Zeichen haben und von einem Bindestrich gefolgt sind: ([aio]-)?(\w{3,}[aio])- Die Identifizierung von Termini erfolgt mittels Suche nach dem Morphemmuster E+W, wobei E ein Präfix und W eine Wortform ist, und mit dem Zeichen „+“ mehrere sukzessive Präfixe gesucht werden können.

Im nächsten Schritt wird eine Gruppierung der extrahierten Termini in zwei Phasen vorgenommen. Zuerst erfolgt die Zuordnung der Termini zu einer Wortfamilie, welche das gleiche Wortelement W enthalten. Dann werden zwei Familien miteinander kombiniert, wenn die Familien Wortformen mit dem gleichen Teilelement enthalten und das gleiche Präfix in einem Terminus jeder Familie vorkommt. Aus der Liste mit Begriffen, die mit Hilfe der Häufigkeitsanalyse extrahiert wurden, wurden zusätzlich diejenigen selektiert, welche in beiden Korpora vorkamen. Log-Likelihood Ratio wurde dazu genutzt, um aus dieser Selektion diejenigen Termini mit einem höheren Wert als 3.8 als Keywords zu filtern. Die beiden Listen mit Termini wurden dann mit einem medizinischen Metathesaurus verglichen.

Die Ergebnisse der Untersuchung zeigten, dass einige fachspezifische Termini mit der angewandten Suche nach Morphem-Mustern nicht identifiziert wurden, sich aber auf der Liste mit Termini befanden, die mittels statistischer Verfahren ermittelt wurden. Umgekehrt wurden fachspezifische Termini korrekt durch die linguistische Methode identifiziert, die mittels statistischer Häufigkeitsanalyse und Log-Likelihood Ratio nicht gefunden wurden. Die Erkenntnis aus dieser Studie lautet daher, dass die besten Ergebnisse durch den Einsatz von linguistischen Methoden zur Terminologie-Extraktion ergänzt durch statistische Methoden zu erwarten sind, was für die Vorteile der hybriden Verfahren spricht.

Einer der meist zitierten hybriden Ansätze stammt von Justeson und Katz (1995, S. 9–27). Der Ansatz geht von der Annahme aus, dass eine hohe Frequenz von Nominalphrasen und die Struktur der Mehrworttermini, die meistens aus maximal vier Wortformen bestehen und aus Nomina und Adjektiven gebildet werden, bereits bestimmen, dass diese Phrasen potentielle Termini sind. Dort wird auch

eine Eigenschaft der terminologischen Mehrworttermini hervorgehoben: Sie sind lexikalisch. Da sich ihre Bedeutung nicht einfach aus der Bedeutung der Wortelemente, aus denen sie sich zusammensetzen, ableiten lässt, ist ihre Aufnahme in ein Wörterbuch erforderlich. Die Konsequenz daraus ist, dass die lexikalischen Mehrworttermini kaum formale Variationen erlauben, da ihre Form im Wörterbuch festgelegt ist. Das sind wiederum Kriterien für die Fachspezifik der Termini, für *termhood*.

Der von Justeson und Katz entwickelte Algorithmus beachtet bei der Analyse der Texte zwei Bedingungen: Die potentiellen Mehrworttermini haben die Mindestfrequenz 2 und bestehen aus Nomina, Adjektiven und einer Präposition. Die Identifizierung von potentiellen Mehrworttermini mit einer bestimmten syntaktischen Struktur erfolgt mittels eines regulären Ausdrucks, mit dem festgelegt wird, dass sie nur eine Präposition enthalten dürfen und mit einem Nomen enden müssen. Die Festlegung der Mindestfrequenz beruht auf der Annahme, dass terminologische Mehrwortphrasen in einem fachspezifischen, technischen Text mehrmals und ohne Variationen wiederholt werden.

Die potentiellen Mehrworttermini wurden mit Hilfe eines fachspezifischen Wörterbuchs auf die Eignung als Termini überprüft. Bei vielen davon konnte kein Eintrag im Fachwörterbuch nachgewiesen werden. Somit bleibt die Entscheidung, ob eine extrahierte Phrase eine terminologische oder eine themenbezogene ist, subjektiv. Dabei muss der sog. „Kontext-Faktor“ berücksichtigt werden. Eine weitere Erkenntnis der Untersuchung betrifft die Qualität der ausgewählten Nominalphrasen: Die niedrige Frequenz der Phrasen in Texten korrelierte mit der Qualität der potentiellen Mehrworttermini: Je niedriger die Frequenz, desto niedriger der Grad von *termhood*. Die festgelegte Mindestfrequenz 2 wirkt umso ungünstiger, je länger der Text ist. Das liegt daran, dass in längeren Abhandlungen auch die nicht terminologischen Mehrwortphrasen entsprechend oft nicht variiert wiederholt werden. Das erfordert eine dynamische Anpassung der Häufigkeitsbedingung abhängig von der Länge der Texte.

Nach der Darstellung von bestehenden Verfahren für die Terminologie-Extraktion wird im folgenden Kapitel die Methode beschrieben, die in dieser Arbeit für die

Extraktion der fachspezifischen Begriffe verwendet wurde. Die Ergebnisse der angewandten Methode, der *Differenzanalyse,* werden im anschließenden Kapitel präsentiert.

5.2 Verwendete Methode bei der Terminologie-Extraktion

Um zu prüfen, inwieweit sich das fachspezifische Vokabular aus den wissenschaftlichen Videos von der verwendeten Fachsprache in den wissenschaftlichen Publikationen unterscheidet, wurden die Fachtermini aus zwei unterschiedlichen Korpora mittels eines statistischen Verfahrens extrahiert. Bevor auf die verwendete Methode eingegangen wird, werden zunächst die Korpora und die vorbereitenden Arbeiten an den Korpora beschrieben, die eine Extraktion von Wortformen und die anschließende Differenzanalyse erst ermöglichen.

5.2.1 Korpora

Es wurden zwei Korpora aus Dokumenten aufgebaut, deren Thematik im Bereich der Energietechnik und Robotik angesiedelt ist. Ein Korpus enthält wissenschaftliche Publikationen, der zweite Korpus besteht aus Transkriptionen von Vorlesungen. Die Sprache aller Dokumente ist Englisch. Die Dokumente wurden durch manuelle Recherchen mittels Suchmaschinen und in Datenbanken gewonnen. Die verwendeten Recherchemethoden und Suchstrategien wurden im Kapitel 3.1 beschrieben.

Das Korpus mit wissenschaftlichen Publikationen umfasst insgesamt 246 Textdokumente, die in zwei Recherchephasen gesammelt wurden: 112 Dokumente stammen aus unterschiedlichen Internet-Quellen sowie Datenbanken und enthalten vollständige Aufsätze im PDF-Format; 134 Dokumente sind nur kurze Abstracts, 115 davon stammen aus der Datenbank *Web of Science*[66] und 18 Dokumente aus der Datenbank *TEMA Technik und Management*[67], welche direkt als Textdokument heruntergeladen werden konnten. Beim Aufbau des Korpus mit

[66] http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/
[67] http://www.wti-frankfurt.de/images/stories/download/de-tema.pdf

wissenschaftlichen Publikationen wurden Aufsätze, technische Berichte sowie Hochschul- und Patentschriften aus Energietechnik und Robotik berücksichtigt.

Das zweite Korpus umfasst die Transkriptionen von 10 Vorlesungen in englischer Sprache zum Thema Energietechnik und Robotik. Sieben davon sind Transkriptionen von Videovorlesungen am Massachusetts Institute of Technology (MIT), die im Internet unter der Lizenz Creative Commons Attribution-Non-Commercial-ShareAlike 3.0 United States veröffentlicht sind. Drei der zehn Transkriptionen geben den Inhalt von öffentlichen Vorlesungen wieder: Zwei davon stammen aus dem Korpus „The British Academic Spoken English“ (BASE) und stehen im Internet für wissenschaftlichen Zwecke frei zur Verfügung, die letzte Transkription stammt von der Australian National University und ist im Internet ohne Angaben zur Nutzungslizenz veröffentlicht. Eine Übersicht mit allen Angaben zu den Vorlesungstranskriptionen ist in der Anlage A1 zu finden. In der Tabelle 5-1 wird die Größe der beiden Korpora gegenübergestellt. In der Spalte *Tokens* wird die Anzahl der einzelnen Bestandteile eines Textes, in *Wortformen* die Anzahl der extrahierten Wortformen angegeben.

Korpus	**Dokumente**	**Tokens**	**Wortformen**
Wissenschaftliche Publikationen	246	920.237	8307
Vorlesungstranskriptionen	10	77.590	1504

Tab. 5-1: Größe der verwendeten Korpora

Für die Durchführung der Differenzanalyse wird ein Referenzkorpus benötigt. Als Referenzkorpus wurde für diese Arbeit eine Liste mit den 5000 häufigsten Wortformen aus dem Corpus of Contemporary American English[68] (COCA) benutzt. Dieses Korpus wurde von Mark Davies[69] erstellt, Professor für Linguistik an der Brigham Young University in Utah.

COCA enthält 450 Millionen Wörter, wird laufend aktualisiert (die letzte Aktualisierung hat im Sommer 2012 stattgefunden) und deckt eine angemessene Auswahl von Gattungen ab. COCA enthält Texte aus Belletristik, Zeitungs- und Zeit-

[68] http://www.wordfrequency.info/top5000.asp
[69] http://davies-linguistics.byu.edu/personal/

schriftenartikeln und wissenschaftlichen Publikationen. Die Texte wurden zwischen 1990 und 2012 gesammelt. Mit einem jährlichen Wachstum von ca. 20 Millionen Wortformen kann COCA als eine Datenbank der modernen englischen Sprache angesehen werden. Die frei verfügbare Liste verzeichnet die häufigsten 5000 Wörter mit Angaben der Frequenz innerhalb des Korpus, der Wortart, der Rangfolge, sowie des Grades der Verteilung im Korpus (dabei steht der Wert 1 für die gleichmäßige Verteilung und der Wert 0 für Wortformen, die sporadisch in wenigen Texten auftreten). Die Tabelle 5-2 zeigt beispielhaft die ersten 10 der häufigsten Wörter im COCA.

Rank	Word	Part of speech	Frequency	Dispersion
1	the	a	22038615	0,98
2	be	v	12545825	0,97
3	and	c	10741073	0,99
4	of	i	10343885	0,97
5	a	a	10144200	0,98
6	in	i	6996437	0,98
7	to	t	6332195	0,98
8	have	v	4303955	0,97
9	to	i	3856916	0,99
10	it	p	3872477	0,96

Tab. 5-2: Die 10 häufigsten Wortformen im COCA

5.2.2 Vorverarbeitung von Texten

Um eine statistische Analyse des Dokumentenkorpus durchführen zu können, müssen die Texte, die meist in unterschiedlichen Formaten vorliegen, so vorbereitet werden, dass die maschinelle Verarbeitung möglich ist. Im ersten Schritt wird eine Formatnormalisierung aller Dokumente vorgenommen (vgl. Granitzer, 2006, S. 439). Es folgt die lexikalische Analyse, während der ein Text in Sätze und Wörter segmentiert wird. In der nächsten Verarbeitungsphase wird eine Merkmalsanalyse durchgeführt, die sich weiter in Lemmatisierung und Part-of-Speech-Tagging unterteilt. Die Merkmalsanalyse bezweckt die Extraktion von Wörtern mit Merkmalen, die für die folgenden Bearbeitungsschritte relevant sind.

Bei der Lemmatisierung werden die Merkmale auf ihren gemeinsamen Merkmalstyp zurückgeführt. Im Fall von Texten sind es die einzelnen Wörter, welche bei der Lemmatisierung auf ihre Grundform, d.h. auf ihren Stamm, zurückgeführt werden. Lemmatisierung ist auch unter den Begriffen Grundformre-

duktion oder Stemming bekannt (vgl. Granitzer, 2006, S. 440; Witschel, 2004, S. 52–53). Beim POS-Tagging wird jedem ermittelten Merkmal die entsprechende Wortformart zugeordnet, zum Beispiel Substantiv, Verb, Adverb, Adjektiv (vgl. Kap. 5.1.2).

Die Formatnormalisierung

Die Formatnormalisierung stellte für die vorliegende Aufgabenstellung zur Terminologie-Extraktion den ersten Schritt dar. Jedes formatierte Dokument enthält neben reinen Text Layout-Informationen (Überschriften, Absatzstrukturen, Seiten-angaben) sowie Metadaten, wie z.B. Autoren und Erstellungsdatum, auch Publikationsinformationen, die keine Bedeutung für die weitere Sprachanalyse haben. Die Hälfte der für die Extraktion vorgesehenen Texte lag im PDF-Format vor. Sie mussten zunächst in ein für die Verarbeitung geeignetes Format konvertiert werden. Alle Texte wurden ins ASCII-Format mit der Kodierung Unicode UTF-8 umgewandelt. Hierfür wurde ein automatisierter Arbeitsablauf *PDF-to-Text* mit der Software *Automator*[70] erstellt. Dieses Programm extrahierte den Text vollständig, entfernte einige Layout-Informationen sowie Formeln, Abbildungen und Tabellen. Folgende Probleme sind aufgetreten:

- die Metadaten wurden in die ASCII-Dokumente mit übernommen. Die Textdokumente enthalten deshalb viele Namen sowie die Institutsangehörigkeit und je nach Publikation sogar Kontaktdaten. Zusätzliche Metadaten in Form von sich wiederholenden Angaben zu Publikation, Jahrgang und Seiten wurden übernommen.
- Der Inhalt von Tabellen wurde extrahiert und als Zahlenreihe in den Textinhalt eingefügt.
- Die Schriftzeichen aus den Formeln wurden extrahiert, die Sonderzeichen wurden aber nur teilweise aufgelöst, teilweise wurden andere Sonderzeichen in das Textdokument eingefügt.

[70] Automator ist ein Programm von Apple, das standardmäßig auf Mac OS X installiert wird. Mit Hilfe von Automator können Routineaufgabe automatisiert durchgeführt werden. Das Programm „PDF to Text“ ist eine der vorinstallierten Standardaktionen in Automator
(vgl. http://www.macosxautomation.com/automator/index.html).

- In vielen Fällen wurde das Wortende nicht korrekt erkannt und mehrere Wörter in einer Einheit zusammengeführt, so dass dadurch Wörter generiert wurden, die in der Form nicht existieren.

Da keine Möglichkeit gefunden wurde, diese Probleme mit Hilfe eines automatisierten Verfahrens zu beseitigen, wurden die Textdokumente für die weitere Bearbeitung so belassen. Auf solche Probleme wird man auch in der Literatur hingewiesen (vgl. Heyer et al., 2006, S. 57–58). Generell werden Probeläufe mit dem Textkonverter empfohlen sowie die Optimierung der Einstellungen und die Aufhebung der Originaldokumente für Prüfzwecke.

Zwei weitere problematische Aspekte, die im Zusammenhang mit der Konvertierung aufgetreten sind, wurden manuell gelöst. Zum einen wurde die harte Silbentrennung, die aufgrund der Formatierung in den Originaldokumenten entstanden ist, unverändert an die Textdokumente weitergegeben. Die harte Silbertrennung wurde mit Hilfe der Funktion *Suchen-und-Ersetzen* aus allen Texten entfernt. Der zweite Aspekt betrifft die Literaturverzeichnisse in den wissenschaftlichen Publikationen. Um zu verhindern, dass viele Namen und Titel sowie andere publikationsbezogene Daten bei der statistischen Analyse mit ausgewertet werden, wurden diese aus allen Texten entfernt.

Die nächsten Verarbeitungsphasen bestehen aus der lexikalischen und Merkmalsanalyse. Wort- und Satzsegmentierung, Lemmatisierung und Part-of-Speech-Tagging wurden mit der Open Source Software GATE[71] durchgeführt.

GATE: die Software für die Textanalyse und Terminologie-Extraktion

Das Akronym GATE steht für General Architecture for Text Engineering. Die Software wurde an der University of Sheffield entwickelt und steht zurzeit in Version 7.0 zur Verfügung. GATE ist eine Framework-Architektur mit drei Typen von Komponenten:

- Die Sprachressourcen (*Language Resources*), welche aus einzelnen Textdokumenten, Dokumentenkorpora, Lexika oder Ontologien bestehen können.

[71] http://gate.ac.uk

- Die Verarbeitungsressourcen (*Processing Resources*), welche die Programme für die Text- und Sprachanalyse darstellen.
- Visuelle Ressourcen (*Visual Resources*), welche die Benutzeroberfläche mit den Funktionen und Tools darstellen, die die Nutzung der beiden vorangestellten Ressourcen erst ermöglichen.

Alle in GATE vorhandenen Ressourcen sind unter der Bezeichnung CREOLE (*Collection of REusable Objects for Language Engineering*) bekannt, welche von der Framework verwendet wird, um die einzelnen benötigten Ressourcen zu finden und hochzuladen. Aus der Sammlung der vorhandenen Ressourcen können je nach Aufgabenstellung einzelne ausgewählt und in das Modul *Applications* geladen werden.

In der Komponente *Applications* hat der Nutzer die Möglichkeit, ein individuelles Programm aus den geladenen Ressourcen zu erstellen. Hier können die Reihenfolge, wie die ausgewählten Programme miteinander kombiniert werden, die Art der Durchführung der in den Algorithmen gespeicherten Verarbeitungsanweisungen und deren Parameter festgelegt werden. So kann z.B. entschieden werden, ob die Algorithmen die Verarbeitung der Dokumente als *pipeline* (d.h. in einer logischen Ordnung von aufeinander folgenden Schritten) oder *parallel, konditional (*d.h. bei der Durchführung muss eine *wenn*-Bedingung erfüllt sein) oder *iterativ* (d.h. die Verarbeitung wird nach einer bestimmten Regel wiederholt) vornehmen. Die analysierte Sprachressource und die Ergebnisse können im *document viewer/editor* eingesehen werden (vgl. Bontcheva, Tablan, Maynard & Cunningham, 2004, S. 352–356).

Die Abbildung 5-2 veranschaulicht die vorangestellte Schilderung. Sie zeigt in der mittleren Spalte ein analysiertes Textdokument, bei dem das Merkmal *Token* aktiviert wurde. Das Merkmal wird im Dokument grün hervorgehoben. Unter dem Dokumentenfenster werden alle im Text gefundenen Merkmale des Typs *Token* aufgeführt, zusammen mit weiteren Angaben, wie z.B. Wortformart und Länge. Nachfolgend werden die für die vorliegende Aufgabenstellung verwendeten GATE-Ressourcen überblickartig dargestellt.

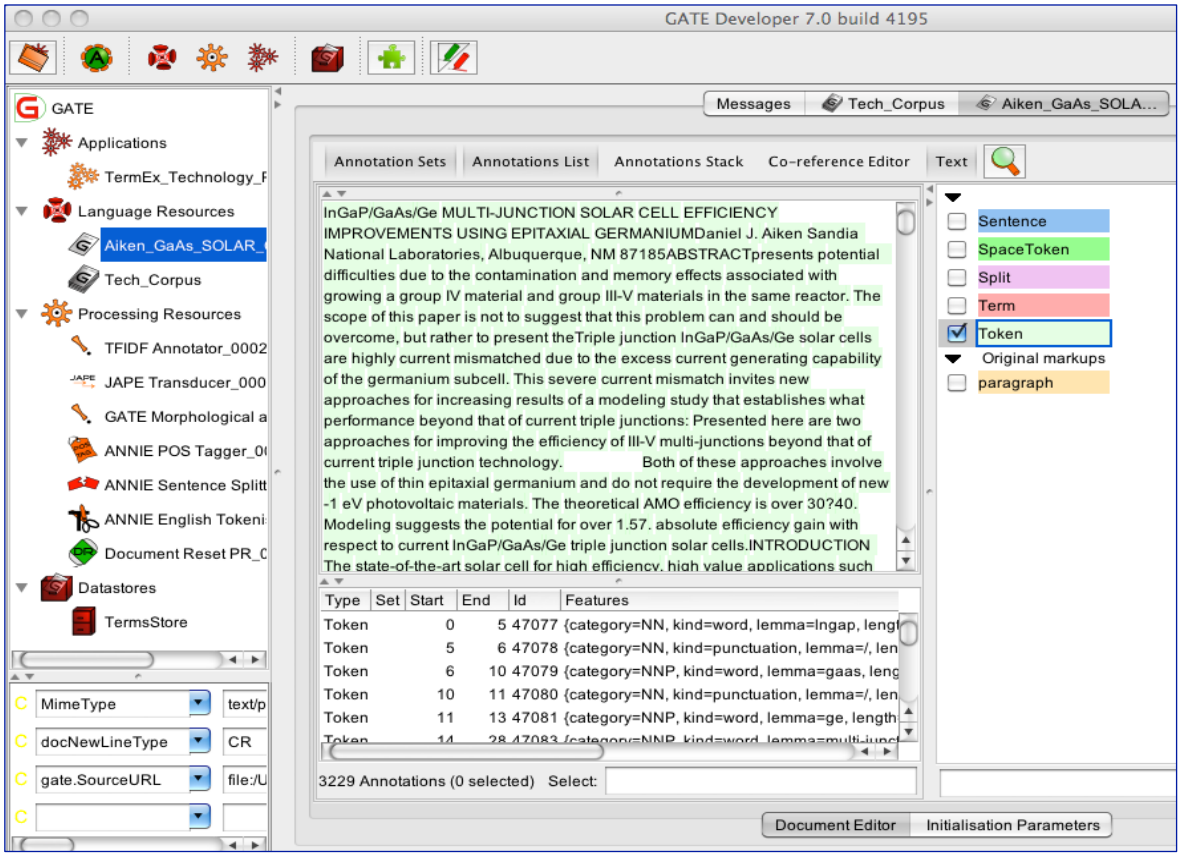


Abb. 5-2: GATE document viewer

Für die Erstellung eines Korpus wurde die Komponente *Language Resources* verwendet. Alle vorbereiteten Textdokumente wurden in den Korpus mit Hilfe der Funktion *Populate* geladen. Der gesamte Korpus wurde anschließend in einer neu eingerichteten *Datastore* in GATE gespeichert. Diese Datenbank-Funktion hat den Vorteil, dass nicht alle Dokumente im Korpus gleichzeitig geöffnet für die Verarbeitung durch die gewählten Algorithmen zur Verfügung gestellt werden, sondern nur ein einzelnes geöffnet, analysiert, erneut geschlossen und in der Datenbank gespeichert wird. Der nächste Schritt bestand in der Auswahl von vier Plugins für die Sprachanalyse (ANNIE, Morphological Analyser, JAPE Transducer, TFIDF Annotator). In *Applications* wurde eine *Corpus Pipeline* eingerichtet, die serielle Verarbeitungsschritte festlegte. Die Abbildung 5-3 zeigt die Reihenfolge der Plugins in der Corpus Pipeline. Im folgenden Abschnitt werden die lexikalische und die Merkmalsanalyse (Wort- und Satzsegmentierung, Lemmatisierung sowie Part-of-Speech Tagging) des Dokumentenkorpus geschildert und die Rolle der ausgewählten Plugins bei der Analyse erklärt.

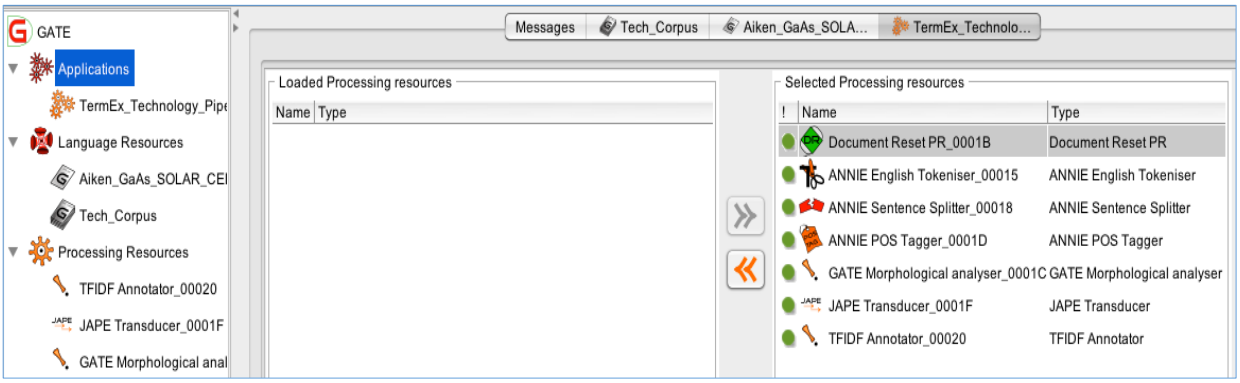


Abb. 5-3: Corpus Pipeline in GATE

Die lexikalische Analyse (Tokenizing)

Um eine Analyse der Texte zu ermöglichen, müssen sie in kleinere Einheiten zerlegt werden. Ein typisches Verfahren ist die Zerlegung von Texten zuerst in Sätze und anschließend die Sätze in Wörter. Diese kleinen Einheiten werden *Tokens* genannt. In der Regel wird dieser Begriff für ein Merkmal vom Typ *Wort* verwendet, aber auch eine Reihenfolge von Zeichen, eine Phrase, eine URL oder eine Email-Adresse sowie ein Akronym können als Token betrachtet werden. (vgl. Konchady, 2008, S. 22). Nach Konchadys (2008, S. 49) Definition „ [...] a token is a string of contiguous alphanumeric characters that may include hyphens and single quotes with a space character on either side“ werden zwei Charakteristika deutlich: Das Merkmal *Token* ist ein alphanumerischer String, der durch Leerzeichen von vorangegangenen und nachfolgenden Strings getrennt ist.
Granitzer (2006, S. 439) beschreibt zwei übliche Methoden für die Segmentierung von Texten:

- nach Wörtern (*Word-Grams*): jedes Merkmal entspricht somit einem Wort,
- nach Charakter *n-Grams*: ein Merkmal entspricht einer Zeichenkette der Länge *n.*

Für die Satz- und Wortsegmentierung der englischen Texte im vorliegenden Korpus wurden zwei Teilkomponenten von ANNIE verwendet: English Tokeniser und Sentence Splitter. Die in ANNIE vorhandenen Typen von Tokens sind: Wörter, Zahlen, Symbole (z.B. $, &, ∧, ∑) sowie Satz- und Leerzeichen. Der English Tokeniser führt in der englischen Sprache spezifische Konstrukte (z.B. „’ll“, „’s“,

„'d“, „'m“, „'til“) in einem Token zusammen und hat darüber hinaus die Aufgabe, negative Ausdrücke - wie z.B. „don't“ - in zwei Tokens zu teilen („do“ und „n't“).

In der Literatur gibt es viele Hinweise auf die Probleme, die im Zusammenhang mit der Segmentierung auftreten können (vgl. Granitzer, 2006, S. 439–440; Witschel, 2004, S. 52–54; Konchady, 2008, S. 49–54). Einige Ursachen, die Segmentierung erschweren können, werden hier beispielhaft genannt:

- Abkürzungen, die von einem Punkt gefolgt sind, werden als Satzschluss interpretiert,
- Fehlende Satz- und Leerzeichen in Folge z.B. einer fehlerhaften Textextraktion,
- Fehlende oder fehlerhafte Silbertrennungszeichen bzw. Bindestriche zwischen den Teilen eines Kompositums.

Die Abbildung 5-4 veranschaulicht die fehlerhafte Satzsegmentierung in einem für diese Arbeit verwendeten Textdokument.

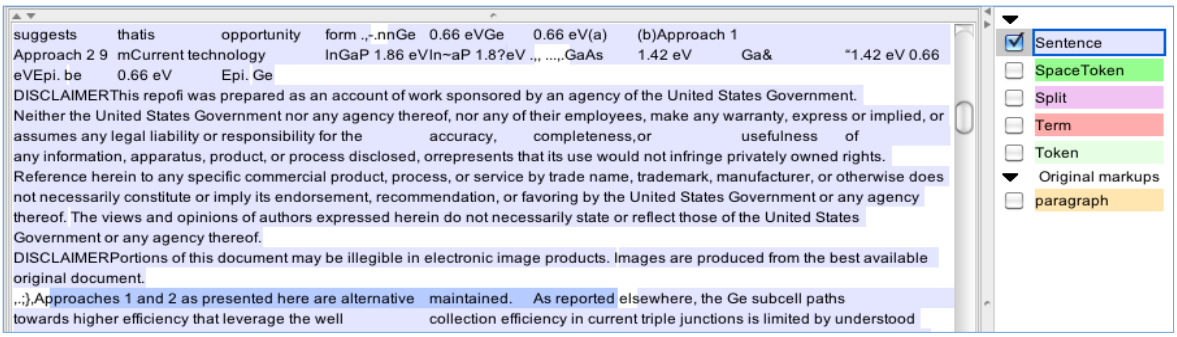


Abb. 5-4: Fehlerhafte Satzsegmentierung in GATE

In der folgenden Abbildung 5-5 wird die fehlerhafte Wortsegmentierung sichtbar, wegen eines Bindestriches mit einem Leerzeichen davor wurden die Bestandteile des Kompositums *hydro-phobic* getrennt.

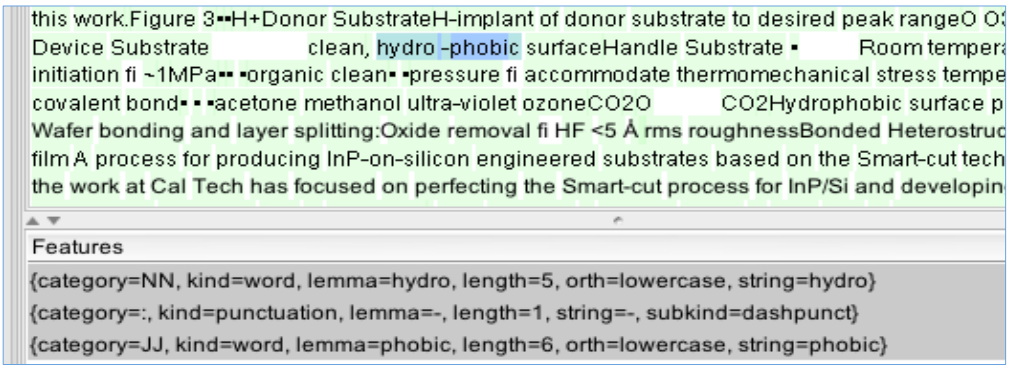


Abb. 5-5: Fehlerhafte Wortsegmentierung in GATE

Die Merkmalsanalyse: Lemmatisierung und Part-of-Speech-Tagging

Bei der Merkmalsanalyse werden die weiteren Eigenschaften von Tokens erfasst und extrahiert. Gerade im Kontext der Terminologie-Extraktion sind nicht die einzelnen in den Texten vorkommenden Wortformen relevant, sondern deren Grundformen. Die Wortformen werden auf ihre Grundform zurückgeführt, indem die Flexionsformen ausgeschaltet werden. Nach Granitzer (2006 S. 440–441) gibt es zwei unterschiedliche Verfahren für die Lemmatisierung: Stemming und Stripping. Beim Stripping werden die Vor- und Nachsilben eines Wortes abgeschnitten. Der Nachteil dieses Verfahrens besteht in der fehlenden Möglichkeit, die irregulären Formen, wie z.B. *sein, bin, war, gewesen* zur Grundform des Wortes zurückzuführen. Im Gegensatz dazu ermöglicht das Stemming-Verfahren die Auflösung von irregulären Formen durch den Einsatz von Wörterbüchern.

Die Grundformreduktion bei Wörtern der englischen Sprache kann aufgrund der wenigen Flexionssuffixe (*-s, -ed, -ing*) relativ leicht durch Abschneiden der Flexive durchgeführt werden. Natürlich gibt es auch in der englischen Sprache Ausnahmen, die beachtet werden müssen. Ein Beispiel: bei der Wortform *sing* muss das Analyseprogramm aufgrund von hinterlegten Regeln erkennen, dass es sich um einen Verbstamm ohne Affix handelt. Aber auch die Wortformen, die mit *-s* enden, bedürfen besonderer Beachtung: Enden die Wörter auf *-ss* oder *-ies,* so darf das letzte *s* nicht abgeschnitten bzw. muss der Flexionssuffix *-ies* abgeschnitten und durch ein *y* ersetzt werden (vgl. Witschel, 2004, S. 54). In GATE übernimmt die Aufgabe der Lemmatisierung der Morphological Analyser, der ein Dokument analysiert, das zuvor in Wörter (Tokens) segmentiert wurde und deren Token-Merkmalen POS-Tags zugewiesen wurden.

Die grammatikalische Analyse und Zuweisung von POS-Tags wurde in GATE automatisch mit dem Plugin ANNIE POS-Tagger durchgeführt. Wie im Kapitel 5.1.2 dargestellt, wird beim POS-Tagging jedes Satzelement mit einer Annotation zur entsprechenden Wortart ergänzt. Der POS-Tagger greift dabei auf einen Regelsatz und ein Lexikon zu, das eine Liste mit Begriffen und die dazugehörigen Wortart-informationen enthält. Konchady (2008, S. 208) weist darauf hin, dass das Lexikon zwischen Klein- und Großschreibung unterscheidet und damit die Zuwei-

sung von korrekten Wortartinformationen ermöglicht. Zum Beispiel wird die Wortform *works* mit dem POS-Tag *Verb* annotiert, die Wortform *Works* mit dem POS-Tag *Substantiv*. In der Regel werden die Substantive in der englischen Sprache kleingeschrieben, es sei denn, sie stehen am Satzanfang oder sind Bestandteile von speziellen Bezeichnungen (sog. *proper name* oder *proper noun*).

Mit den zwei letzten Plugins *JAPE Transducer* und *TFIDF Annotator* wird das Programm zur Extraktion von Wortformen aus dem annotierten Dokumentenkorpus vervollständigt. Das Akronym *JAPE* wurde von Java Annotation Patterns Engine gebildet. Der JAPE Transducer konvertiert die Annotationen in einem Dokument in Sequenzen, die mit Hilfe von regulären Ausdrücken gefunden werden können. Innerhalb des GATE-Programms besteht die Aufgabe des JAPE Transducer darin, alle Annotationen eines bestimmten Typs zu filtern. Für die vorliegende Aufgabenstellung werden alle Wortstämme benötigt. Der Transducer wird eingesetzt, um alle Merkmale mit der Annotation *Token* zu filtern, wobei die Tokens vom Typ *Wort* sind und eine Mindestlänge von vier Zeichen haben. Weiterhin soll der JAPE Transducer von den gefundenen Tokens nur den Wortstamm zurückliefern. Der TFIDF Annotator ist ein externer Plugin, welcher die Aufgabe hat, die Häufigkeit der Wortformen innerhalb des Korpus sowie die Anzahl der Dokumente zurückzugeben, in denen die extrahierten Wortformen vorkommen.

In den vorangestellten Darstellungen wurde gezeigt, welche Voraussetzungen erfüllt sein müssen, um aus einem Textkorpus die statistischen Informationen zu entnehmen, die Aufschluss über das Vorhandensein von Terminologie geben können. In der Phase der Formatnormalisierung werden Layout-Informationen und eventuell vorhandene, aber nicht benötigte Metadaten entfernt und alle Texte des Korpus in ein bestimmtes Format umgewandelt. Es folgt eine lexikalische Analyse mit Satz- und Wortsegmentierung. Die grammatikalische Analyse der Texte schließt dort an: Die Wortformen werden mit Informationen zur Wortart versehen. Die letzte Phase der Vorverarbeitung besteht in der Lemmatisierung der Wortformen, sodass eine Analyse der Wortstämme möglich wird. Das verwendete statistische Verfahren trägt die Bezeichnung Differenzanalyse und wird im nächsten Abschnitt beschrieben.

5.2.3 Terminologie-Extraktion mittels Differenzanalyse

Im Kapitel 5.1.1 *Statistische Methoden* wurde die grundsätzliche Vorgehensweise bei der Durchführung der Differenzanalyse geschildert. Voraussetzung ist die Bereitstellung von zwei Korpora. Das Analysekorpus wird mit Texten einer bestimmten Fachrichtung gebildet. Das zweite Korpus, Referenzkorpus genannt, besteht aus Texten in Allgemeinsprache. Für die Wortformen oder Wortformkombinationen der beiden Korpora werden die Häufigkeitswerte ermittelt und die Mengendifferenz der fachspezifischen und der allgemeinsprachlichen Wortformen berechnet.

Ziel der vorliegenden Aufgabenstellung war es herauszufinden, wie sich die Mengen des verwendeten Fachvokabulars in den beiden Korpora unterscheiden. Es wurde die Häufigkeitsverteilung der Wortformen in den im Kapitel 5.2.1 vorgestellten Korpora verglichen.

Mit Hilfe von GATE wurden aus den Korpora Einwortbegriffe und diejenigen Komposita extrahiert, welche aus zwei oder drei Wortformen bestehen und mit Bindestrich verbunden sind: 8307 Wörter aus dem Korpus mit wissenschaftlichen Publikationen und 1504 aus dem Korpus mit Vorlesungstranskriptionen. Diese Wortformen stellen die Wortstämme dar (vgl. Kap. 5.2.1). Zusammen mit den extrahierten Wortstämmen lieferte das Programm die Anzahl jedes Begriffs innerhalb des Korpus und die Anzahl der Dokumente innerhalb des Korpus, in denen der jeweilige Begriff vorkommt.

Mit der Funktion *SVerweis* der Software Microsoft Excel wurde anschließend überprüft, welche und wie viele der extrahierten Begriffe aus den Analysekorpora im Referenzkorpus COCA vorkommen. Mit Hilfe der *Filterfunktion* in Excel wurden die Termini gefiltert, die in einem der Analysekorpora vorkamen, aber im Referenzkorpus nicht nachgewiesen werden konnten. Die Ergebnisse dieser Analyse wurden in zwei Wortformklassen eingeteilt:

1. Klasse der Fachausdrücke, welche alle Wortformen enthält, die nur in den Analysekorpora vorkommen, da sie sehr wahrscheinlich Fachtermini sind.
2. Klasse der Wortformen, die sowohl in den Analysekorpora als auch im Refe-

renzkorpus vorkommen, aber mit unterschiedlichen Häufigkeiten. Um weitere Fachtermini zu identifizieren, sollen aus dieser Klasse die Wortformen mit einer deutlich höheren Häufigkeit als im Referenzkorpus ermittelt werden.

Bei diesen Wortformen, die in den Analysekorpora häufiger als im Referenzkorpus auftreten, kann die Frequenz nicht das einzige Kriterium bei der Identifizierung von Fachtermini darstellen. Deshalb wurde eine statistische Prüfgröße als Schwellenwert hinzugezogen. Als statistische Prüfgröße wurde in Anlehnung an Witschel (2004, S. 57–61) der Häufigkeitsquotient gewählt. Nach Witschel lässt sich dieses Maß berechnen, indem die relative Häufigkeit des Worts *w* im Fachtext durch seine relative Häufigkeit im Referenzkorpus geteilt wird. Das wird mit der folgenden Formel ausgedrückt:

$$hq(w) = \frac{k}{n}\frac{1}{p}$$

Mit *k* wird die Frequenz des Wortes *w* im Analysekorpus bezeichnet, *n* gibt die Länge des Analysekorpus in Wörtern an und *p* ist die relative Häufigkeit des Wortes *w* im Referenzkorpus. Um die relative Häufigkeit der aus den Analysekorpora extrahierten Wortformen im Referenzkorpus berechnen zu können, wurde die auf der Webseite von COCA angegebene Korpusgröße von 450 Millionen Wörtern verwendet. Anschließend wurde die Liste der Termini nach dem Wert des Häufigkeitsquotienten absteigend sortiert und festgelegt, dass Wortformen mit einer bestimmten Mindestfrequenz und einem Häufigkeitsquotienten über 10 einer zweiten Gruppe von Termini zugeordnet werden, die mit einer gewissen Wahrscheinlichkeit fachspezifische Ausdrücke sind. Ein Fachexperte könnte dann anhand der Frequenz der Termini im fachspezifischen Korpus, des Wertes der Prüfgröße für die jeweiligen Termini aus der zweiten Gruppe und einiger Beispielsätze, welche den Begriff enthalten, entscheiden, ob diese Wörter den Status „Terminus“ erhalten sollten. Witschel (2004, S. 32–33; S. 63) verwendet diese aus dem Bereich des Information Retrieval stammende Methode mit der Bezeichnung *Relevance Feedback*, um die Recall- und Precision-Werte zu steigern.

5.3 Darstellung der Ergebnisse

Aus dem Korpus mit wissenschaftlichen Publikationen (*WissPubl*) wurden 8307 Wörter und aus dem Korpus mit Transkriptionen von Vorlesungen (*TranskriptVorl*) 1504 Wörter extrahiert. Als Referenzkorpus (*COCA*) wurde eine Liste mit 5000 Wörter verwendet, die in der allgemeinen Sprache am häufigsten vorkommen.

5.3.1 Ergebnisse der Analyse des Korpus WissPubl

5364 Wörter (64,57% der Gesamtanzahl der Wortformen) aus dem Korpus WissPubl kamen im COCA nicht vor. Diese Wörter werden der ersten Klasse der Fachausdrücke zugeordnet, die nur diejenigen Wortformen enthält, welche im Analyse-, aber nicht im Referenzkorpus vorkommen und als Fachtermini gelten. Diese erste Klasse wird im Folgenden *Klasse der Fachtermini* genannt. Von den 8307 Wörter aus dem Korpus WissPubl waren 2943 Wörter (das sind 35,42%) auch im COCA nachgewiesen. Von der Gesamtanzahl der extrahierten Wortformen aus dem Korpus WissPubl sind 733 Komposita (das entspricht 8,82% von 8307 Wortformen) und 7574 (das sind 91,17%) Einworttermini. In der Tabelle 5-3 werden diese ersten Ergebnisse aus den beiden Analysekorpora gegenübergestellt.

	Korpus Wissenschaftliche Publikationen *WissPubl*	Korpus Transkriptionen Vorlesungen *TranskriptVorl*	Referenzkorpus *COCA*
Anzahl Wortformen	8307	1504	5000
Anzahl Wortformen, die nur im Analysekorpus vorkommen (= 1. Klasse der Fachtermini)	5364	264	
Anteil vom Gesamtkorpus	64,57%	17,55%	
Anzahl Wortformen aus dem Analysekorpus, die mit dem Referenzkorpus übereinstimmen	2943	1240	
Anteil vom Gesamtkorpus	35,42%	82,44%	
Anzahl von Komposita	733	6	
Anteil vom Gesamtkorpus	8,82%	0,40%	
Anzahl von Einworttermini	7574	1498	
Anteil vom Gesamtkorpus	91,17%	99,60%	

Tab. 5-3: Analyseergebnisse der extrahierten Wortformen aus WissPubl und TranskriptVorl

Von den 5364 Wörtern aus der Klasse der Fachtermini haben 23 Wörter eine Häufigkeit zwischen 400 und 1421. Nur zwei Wörter von den 23 weisen eine absolute Häufigkeit von mehr als 1400 auf, das sind *electrode* und *datum*. Eine Übersicht mit den häufigsten 25 Wörtern aus den Analysekorpora gibt Tabelle 5-4.

WissPubl **1. Klasse der Fachtermini**				**TranskriptVorl** **1. Klasse der Fachtermini**		
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Terminus	Dokument-frequenz	Wort-frequenz
1	electrode	36	1421	capacitor	3	116
2	*datum*	127	1401	*voltage*	8	88
3	lithium	35	854	volt	5	79
4	simulation	64	819	zero	8	78
5	duct	18	721	minus	8	59
6	*thermal*	93	701	watt	4	58
7	parameter	74	696	*thermal*	3	52
8	renewable	58	656	joule	3	50
9	*voltage*	51	603	resistor	3	44
10	electrolyte	29	598	capacitance	3	41
11	*discharge*	32	593	inverter	2	40
12	*higher*	99	575	*discharge*	5	37
13	silicon	16	573	dissipate	4	37
14	configuration	53	560	metre	3	36
15	ventilation	30	533	*higher*	8	29
16	interface	53	522	ion	2	28
17	electrochemical	30	487	percent	3	28
18	leakage	17	460	multiply	7	27
19	savings	56	457	radius	4	27
20	controller	40	452	ampere	2	26
21	algorithm	46	413	thickness	2	26
22	robotic	28	409	*datum*	3	25
23	further	99	403	copper	3	24
24	heating	70	388	uranium	3	23
25	plenum	9	386	kilometer	3	21

Tab. 5-4: Die 25 häufigsten Wörter aus den Analysekorpora

Die manuelle Überprüfung der Begriffe aus der Klasse der Fachtermini in beiden Analysekorpora ergab, dass ein Anteil der mittels Differenzanalyse ermittelten Wortformen aus Eigennamen (*Named Entities),* Adjektiven, Zahlen, fremdsprachigen Wörtern und anderen Begriffen besteht, wie sie üblicherweise in den Metadaten der Publikationen vorkommen. Im Korpus WissPubl wurde ein Anteil von

6,41% (344 Wörter) von den 5364 Wortformen dieser Klasse gefunden. Diese Wortformen stellen keine fachspezifischen Termini dar und wurden daher nicht näher betrachtet. In der Tabelle 5-5 wurden die Anteile der nicht fachspezifischen Wortformen in der Klasse der Fachtermini gegenübergestellt.

	Korpus *WissPubl*	**Korpus *TranskriptVorl***
Anzahl Wortformen	8307	1504
Anzahl Wortformen, die nur im Analysekorpus vorkommen (1. Wortformklasse = Klasse der Fachtermini)	5364	264
Anteil vom Gesamtkorpus	64,57%	17,55%
Anzahl der in der Klasse von Fachtermini enthaltenen nicht fachspezifischen Begriffe	344	14
Anteil von nicht fachspezifischen Begriffen in der Klasse von potentiellen Fachtermini	6,41%	5,30%
Endgültige Wortformenanzahl in der Klasse der Fachtermini nach Abzug der nicht fachspezifischen Begriffe	5020	250

Tab. 5-5: Anzahl von nicht fachspezifischen Wörtern in der Klasse der Fachtermini der Analysekorpora

Die in der Klasse der Termini ermittelten nicht fachspezifischen Wortformen wurden folgenden Kategorien zugeordnet: Adjektive, geographische Namen, Monats- und Personennamen, Institutionen, Soft- und Hardware, fremdsprachige Begriffe, Zahlen und Ordinalzahlen, Wochentage und Währungen. Eine tabellarische Übersicht über die daraus gebildeten Kategorien mit deren Wortanzahl und einige Beispiele aus jeder Kategorie gibt die Anlage A3. Die Tabelle 5-6 zeigt die häufigsten fünf Wörter von den 344 Wortformen aus WissPubl, die als nicht fachspezifisch eingestuft wurden, obwohl sie eine hohe Frequenz aufweisen. Die Gesamtheit der nicht fachspezifischen Wörter aus der Klasse von potentiellen Fachtermini beider Korpora ist in den Anlagen A4 und A5 aufgelistet.

Rang	Terminus	Dokument-frequenz	Wort-frequenz
23	further	99	403
35	california	38	301
64	greater	58	200
67	larger	55	195
91	simon	5	158

Tab. 5-6: Die fünf häufigsten nicht fachspezifischen Wörter aus WissPubl

Beide Analysekorpora wurden anschließend mittels Differenzanalyse auf die Anzahl von gemeinsamen Begriffen in der Klasse der Fachtermini untersucht. Die Klasse der Fachtermini von WissPubl enthält nach der Entfernung der nicht fachspezifischen Begriffe noch 5020 Wortformen. 4839 Wörter (96,39%) kommen nur in WissPubl vor (vgl. Tabelle 5-8). 181 Begriffe kommen in beiden Korpora vor. Die Besonderheiten der 181 gemeinsam vorkommenden Begriffe werden im Abschnitt *Gemeinsamkeiten der beiden Korpora* dargestellt.

Die 4839 Wörter, die nur im WissPubl vorkommen, wurden nach Häufigkeit absteigend geordnet und mit Hilfe der Excel-Funktion *Häufigkeiten* in Frequenzklassen eingeteilt. Es wurden 13 Frequenzklassen mit unterschiedlicher Klassenbreite gebildet. Die Klassengrenzen wurden anhand der Daten festgelegt, sodass die Klassen unterschiedlich stark besetzt sind. Eine tabellarische Übersicht über die gebildeten Frequenzklassen mit der Anzahl der enthaltenen Wörter und einigen Beispielbegriffen befindet sich in der Anlage A6. Das Diagramm in der Abbildung 5-7 veranschaulicht die Verteilung der Häufigkeitswerte innerhalb der Klasse von Fachtermini aus WissPubl.

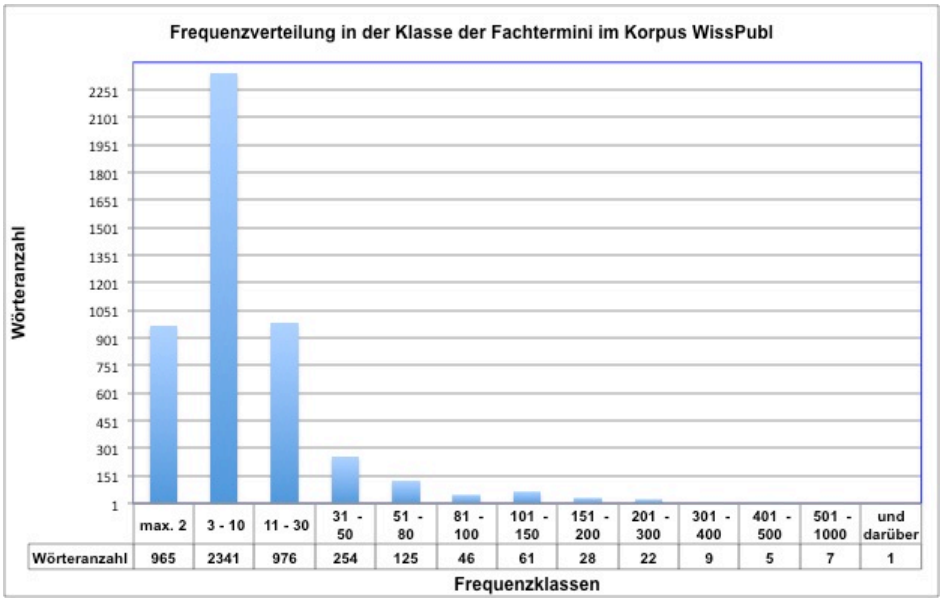


Abb. 5-7: Frequenzklassen der Fachtermini in WissPubl

Die umfangreichste Frequenzklasse enthält 2341 Begriffe, welche zwischen drei und zehnmal im Korpus vorkommen. Einige Beispiele von Nomina aus dieser Klasse sind: *congestion, polymerization, single-junction, desorption, nano-*

structure, multiplexer und *ultrasonic.* Zwei Frequenzklassen enthalten ungefähr die gleiche Anzahl von Termini: die Klasse der Begriffe mit zweimaligem Vorkommen und die Klasse der Begriffe, die zwischen 11 und 30-mal vorkommen. Die Klasse mit Wortformen, die nur zweimal vorkommen, enthält 965 Wörter. Auffallend sind hier neben fachspezifischen Begriffen, wie z.B. *nano-composite, voltage-current, polyphase, interpenetrate,* viele Abkürzungen sowie aufgrund von Konversionsfehlern falsch geschriebene Wörter. Einige Beispiele: *systemreset, elbestawlt, bruedj*, *optimizationstrategy.*

Insbesondere fallen die fünf Frequenzklassen mit einer kleinen Anzahl von Begriffen und hohen Frequenzen auf. Diese Frequenzklassen (151–200 bis 501–1000) enthalten insgesamt 71 Termini. Nur zwei davon sind Adjektive: *optimal* und *porous*. Es wurden außerdem zwei Abkürzungen ermittelt, die sehr häufig verwendet wurden: *hvac* und *http*. Die Termini *electrode* und *datum* haben die mit Abstand höchste Frequenz von 1421 bzw. 1401. *Electrode* wurde in 36 (14,63%), *datum* in 127 (51,62%) von 246 Dokumenten des Korpus WissPubl erwähnt. Nach einer manuellen Überprüfung dieser Begriffe wurde festgestellt, dass die hohe Frequenz der Termini und die niedrige Dokumentenanzahl im Korpus, in denen sie enthalten sind, auf hohe Fachspezifik hinweisen. In den Anlagen A7 bis A7d ist eine Übersicht mit den häufigsten Wortformen jeder Frequenzklasse aus der ersten Klasse der Fachtermini vom Korpus WissPubl. Tabelle 5-7 zeigt ausgewählte Fachtermini aus fünf Frequenzklassen mit den höchsten Häufigkeiten.

Frequenzklasse: 151-300			**Frequenzklasse: 301-1000**		
Terminus	Dokument-frequenz	Wort-frequenz	Terminus	Dokument-frequenz	Wort-frequenz
calibrate	19	154	cathode	22	314
autonomous	23	153	plenum	9	386
blower	11	151	junction	22	343
porosity	9	163	robotic	28	409
lithium-ion	21	179	controller	40	452
polymer	19	190	leakage	17	460
abstraction	8	191	electrochemical	30	487
calibration	10	208	ventilation	30	533
electron	28	206	silicon	16	573
absorption	21	234	electrolyte	29	598
graphene	7	234	parameter	74	696
cells	25	260	duct	18	721
anode	24	286	simulation	64	819
payload	19	298	lithium	35	854

Tab. 5-7: Beispieltermini aus Frequenzklassen mit den höchsten Häufigkeiten

5.3.2 Ergebnisse der Analyse des Korpus TranskriptVorl

264 Wörter (17,55%) aus dem Korpus TranskriptVorl kamen im COCA nicht vor. Diese Wortformen werden der ersten Wortformklasse zugeordnet, die nur diejenigen Wortformen enthält, welche im Analyse- aber nicht im Referenzkorpus vorkommen und als Fachausdrücke gelten. Analog zu WissPubl wird diese erste Wortformklasse auch *Klasse der Fachtermini* genannt. Von den 1504 Wörtern aus TranskriptVorl wurden 82,44% (1240 Wörter) im COCA nachgewiesen. Von 1504 Wortformen in TranskriptVorl sind 99,6% (1498 Wörter) Einworttermini und nur 0,4% Komposita (sechs Wortzusammensetzungen). Eine Gegenüberstellung dieser Ergebnisse mit denen aus WissPubl befindet sich in der Tabelle 5-3. In der Klasse der Fachtermini hat der Terminus *capacitor* die mit Abstand höchste Häufigkeit von 116. Weitere sieben Wörter haben eine Häufigkeit zwischen 50 und 88 (vgl. Tabelle 5-4).

Auch die der Klasse der potentiellen Fachtermini zugeordneten Begriffe aus diesem Korpus wurden zwecks Identifizierung von nicht fachspezifischen Begriffen manuell überprüft. In TranskriptVorl wurde ein Anteil von 5,3% (14 Wörter) von den 264 Wortformen ermittelt, die Eigennamen (*Named Entities*) waren und daher aus der Klasse der potentiellen Fachtermini entfernt wurden. Alle 14 nicht fachspezifischen Wörter sind in der Anlage A5 aufgelistet. Die 14 ermittelten Eigennamen ließen sich folgenden Kategorien zuordnen: geographische Bezeichnungen, Monats- und Personennamen, Wochentage und Währungen. Eine tabellarische Übersicht über die daraus gebildeten Kategorien mit Wörteranzahl und einigen Beispielen ist in der Anlage A3 zu finden. Sechs Eigennamen sind in beiden Korpora vorhanden: *California, Germany, Japan, Peter, February, October*. In der Tabelle 5-5 wurden die Anteile der nicht fachspezifischen Wortformen in der Klasse der potentiellen Fachtermini der beiden Korpora gegenübergestellt.

Die Klasse der Fachtermini der beiden Analysekorpora wurde mittels Differenzanalyse auf die Anzahl von gemeinsamen Begriffen untersucht. Die Klasse der Fachtermini von TranskriptVorl enthält nach der Entfernung der nicht fachspezifischen Begriffe noch 250 Wortformen. 69 Wörter (27,60%) kommen nur in TranskriptVorl vor (vgl. Tabelle 5-8). Diese wurden nicht weiter untersucht.

Im folgenden Kapitel werden die Analyseergebnisse der 181 Wortformen beschrieben, die nach Bildung der Mengendifferenz in den Korpora nachgewiesen wurden.

	Korpus *WissPubl*	**Korpus *TranskriptVorl***
Gesamtanzahl Wortformen	8307	1504
Anzahl Wortformen in der Klasse der Fachtermini nach der ersten Differenzanalyse	5364	264
Anzahl der entfernten, nicht fachspezifischen Begriffe	344	14
Endgültige Wortformenanzahl in der Klasse der Fachtermini	5020	250
Anzahl Wortformen in der Klasse der Fachtermini, die nach der zweiten Differenzanalyse nur im jeweiligen Korpus nachgewiesen wurden	4839	69
Anteil von Fachtermini, die nur in einem Korpus vorkommen	96,39%	27,60%
Anzahl Fachtermini, die in beiden Korpora vorkommen	181	181
Anteil von Fachtermini, die in beiden Korpora vorkommen	3,60%	72,40%

Tab. 5-8: Anzahl von Fachtermini der beiden Korpora nach unterschiedlichen Analysephasen

5.3.3 Gemeinsamkeiten der Korpora WissPubl und TranskriptVorl

Von den 25 häufigsten Wörtern der Klasse der Fachtermini sind fünf in beiden Korpora vorhanden: *datum, voltage, discharge, thermal, higher*. Das Wort *higher* befindet sich aufgrund der hohen Frequenz in der Klasse der Fachtermini, obwohl es kein Fachwort ist. Es kann angenommen werden, dass *higher* überwiegend Bestandteil eines Mehrwortterminus ist (vgl. Anlage A2).

Die erste Klasse der Fachtermini der beiden Analysekorpora wurde mittels Differenzanalyse auf die Menge der gemeinsamen Begriffe untersucht. Von den 5020 Wörtern des Korpus WissPubl sind 181 (3,6%) Wörter in der Klasse der Fachtermini des Korpus TranskriptVorl enthalten. Also sind 72,4% der Fachtermini aus TranskriptVorl in WissPubl enthalten (vgl. Tabelle 5-8). In den Anlagen A9 bis A9b werden die gemeinsam vorkommenden Begriffe in den Analysekorpora mit ihren Frequenzen im jeweiligen Korpus einander gegenübergestellt. Diese 181 Fachtermini wurden anschließend nach Frequenzklassen gruppiert.

Es wurden 14 Frequenzklassen mit unterschiedlichen Klassenbreiten gebildet, analog zum Korpus WissPubl. Die Klassengrenzen wurden anhand der Daten aus WissPubl festgelegt. Eine tabellarische Übersicht über die gebildeten Frequenz-

klassen mit der Anzahl der enthaltenen Wörter und einigen Beispielbegriffen befindet sich in der Anlage A8. Das Diagramm in der Abbildung 5-8 veranschaulicht die Verteilung der Häufigkeitswerte nach Frequenzklassen.

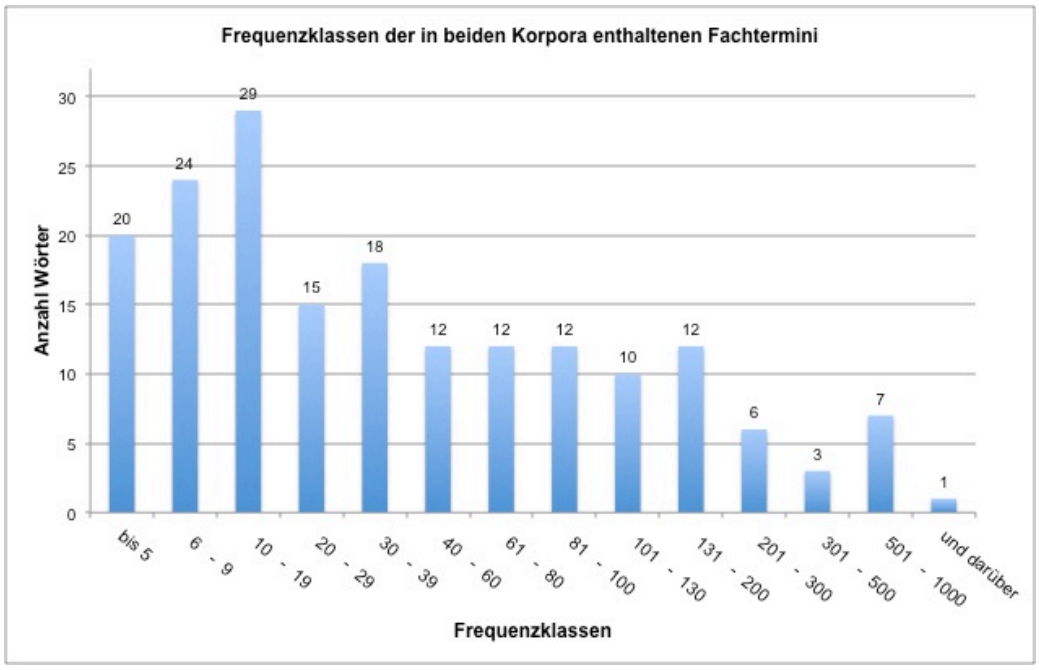


Abb. 5-8: Frequenzklassen der in beiden Korpora enthaltenen Fachtermini

Hier zeigt sich eine gleichmäßige Verteilung der Häufigkeiten im mittleren Bereich. Fünf Frequenzklassen enthalten etwa die gleiche Anzahl von Termini: vier Klassen jeweils mit 12 und die Frequenzklasse 101–130 mit 10 Termini. Die Frequenzklasse 10–19 umfasst 29 Termini (16% von 181 Begriffen) und ist somit die größte Klasse. Etwa gleich stark besetzt sind die Frequenzklassen bis 5, 6–9 und 30–39. Insgesamt enthalten sie 62 Termini (34,25% von 181 Begriffen). Die Anzahl der Wortformen mit einer Frequenz ab 200 ist gering: 17 (9,39% von 181) Termini verteilen sich über vier Häufigkeitsklassen.

Der Terminus *datum* hat die höchste Frequenz und weist somit eine hohe Fachspezifik auf. Er kommt in 127 (51,6% von 246) Dokumenten von WissPubl 1401-mal vor. Die Frequenzklasse 501–1000 enthält nur sieben Termini: *thermal, renewable, voltage, discharge, higher, configuration* und *interface*. *Renewable* ist ein Adjektiv, welches in den ausgewählten Publikationen zum Thema Energietechnik zusammen mit dem Terminus *energy* einen sehr häufig verwendeten

Mehrwortterminus bildet. Auffallend ist die häufige Nutzung des Adjektivs *high*, hier in der Steigerungsform *higher*. Nur drei Nomina enthält die Frequenzklasse 301–500: *algorithm, hybrid* und *thickness*. Von den sechs in der Frequenzklasse 201–300 enthaltenen Begriffen haben *grid* und *zero* die höchste Frequenz mit 297. *Zero* wurde in 58 Dokumenten (23,57%) von WissPubl verwendet und ist Teil des Mehrwortbegriffs *Zero energy home*. In TranskriptVorl wurde *zero* in acht Texten 78-mal verwendet.

Bei den vier Frequenzklassen, die jeweils 12 Termini enthalten, wurden die Häufigkeitswerte in WissPubl mit denen in TranskriptVorl verglichen. Zuerst wurden diejenigen Begriffe betrachtet, die in TranskriptVorl nur zweimal vorkamen. Acht Wörter von 48 Termini dieser vier Klassen sind in WissPubl vergleichsweise häufig*: prototype, matrix, optimized, excess, extract, simplify, foil und efficiently*. Diese Begriffe haben in WissPubl Häufigkeitswerte zwischen 47 (*efficiently*) und 188 (*prototype*). Zwei Begriffe von TranskriptVorl heben sich durch die hohen Häufigkeitswerte im Vergleich mit deren Häufigkeit in WissPubl hervor: *capacitor* (116) und *watt* (58). Der Begriff *inverter* kommt mit etwa gleicher Frequenz in beiden Analysekorpora vor. Neben den Adjektiven *renewable* und *higher,* die sich in den Klassen mit hohen Frequenzen befinden, kommen hier drei weitere vor: *static* (148), *linear* (144) und *smaller* (125).

In der Frequenzklasse 10-19 fallen die Termini *resistor, volt* und *minus* auf. Diese kommen in TranskriptVorl besonders oft vor, verglichen mit WissPubl. Termini wie z.B. *uranium, sulfuric, kilowatt-hour, Fahrenheit, joule, millimeter* und *infinity* sind im Korpus WissPubl eher selten (Frequenzklasse *bis 5*). In dieser Klasse fällt der Begriff *joule* auf, weil er in TranskriptVorl 50-mal vorkommt. In der Tabelle 5-9 werden die Unterschiede in den Häufigkeitswerten einiger gemeinsamen Begriffe in den beiden Korpora veranschaulicht.

	WissPubl		**TranskriptVorl**	
Terminus	Dokument-frequenz	Wort-frequenz	Dokument-frequenz	Wort-frequenz
dissipation	3	9	2	9
convection	5	8	2	15
Dissipate	4	7	4	37
Orbital	4	7	2	11
Joule	2	5	3	50
Uranium	2	3	3	23

Tab. 5-9: Frequenzen der gemeinsam vorkommenden Fachtermini in den Analysekorpora

5.3.4 Ergebnisse der Analyse der zweiten Klasse der Wortformen

Nach der Differenzanalyse wurden die Begriffe der beiden Analysekorpora in Wortformenklassen eingeteilt. Die erste Wortformenklasse enthielt nur die Fachausdrücke, welche in beiden Analysekorpora, aber nicht im Referenzkorpus vorkamen. Diese wurde *Klasse der Fachtermini* genannt. Die Ergebnisse der Untersuchung dieser Klasse wurden bereits dargestellt.

Die zweite Wortformenklasse besteht aus den Begriffen, die sowohl in Analysekorpora als auch im Referenzkorpus vorkommen. Die zweite Wortformenklasse hat in WissPubl einen Umfang von 2943 Begriffen (35,42% vom Gesamtkorpus) und in TranskriptVorl enthält sie 1240 Begriffe (82,44% vom Gesamtkorpus). In der Tabelle 5-10 werden die Wortformenklassen aus den beiden Korpora gegenüber gestellt.

	Korpus *WissPubl*	**Korpus *TranskriptVorl***
Gesamtanzahl Tokens	920.237	77.590
Gesamtanzahl Wortformen	8307	1504
Anzahl Wortformen, die nur im Analysekorpus vorkommen (1. Klasse der Fachtermini)	5020	250
Wortformenanzahl in der Klasse der Fachtermini, die nur im jeweiligen Korpus vorhanden sind	4839	69
Anteil von Fachtermini, die nur in einem Korpus vorkommen	96,39%	27,60%
Anzahl Fachtermini, die in beiden Korpora vorkommen	181	181
Anteil von Fachtermini, die in Analysekorpora vorkommen	3,60%	72,40%
Anzahl Wortformen in der 2. Wortformklasse	2943	1240
Anteil vom Korpus	35,42%	82,44%

Tab. 5-10: Wortformenklassen in den Analysekorpora

In der zweiten Wortformenklasse wurden mit Hilfe des Häufigkeitsquotienten diejenigen Begriffe identifiziert, die im Analysekorpus mit einer deutlich höheren Häufigkeit als im Referenzkorpus vorkommen. Für die Berechnung des Häufigkeitsquotienten wurde in Anlehnung an Witschel (2004, S. 61) die relative Häufigkeit eines Wortes im Analysekorpus durch seine relative Häufigkeit im Referenzkorpus nach der folgenden Formel geteilt: $hq(w) = \frac{k}{n}\frac{1}{p}$ (vgl. Kap 5.2.3).

Zuerst wurde mit Hilfe der Excel-Funktion *SVerweis* die absolute Häufigkeit jedes Wortes aus dem Analysekorpus im Referenzkorpus ermittelt. Danach wurde die relative Häufigkeit des Wortes im Analysekorpus und im Referenzkorpus berech-

net. Die relative Häufigkeit eines Wortes ergibt sich aus der Frequenz des Wortes im Korpus geteilt durch die Gesamtzahl der Wörter des Korpus. Um die relative Häufigkeit der aus den Analysekorpora extrahierten Wortformen zum Referenzkorpus berechnen zu können, wurde die auf der Webseite von COCA angegebene Korpusgröße von 450 Millionen Wörtern verwendet. Anschließend wurde der Häufigkeitsquotient nach der oben angegebenen Formel errechnet. Um potentielle Termini in dieser zweiten Wortformenklasse zu ermitteln, wurden dann nur die Wortformen betrachtet, deren Häufigkeitsquotient größer als 10 ist.

In WissPubl entsprechen diesem Kriterium 130 Wörter (4,42% von 2943) bzw. in TranskriptVorl 57 Wörter (4,60% von 1240) Wörtern in der zweiten Wortformklasse. Werden nun nur die Wörter betrachtet, die eine Mindestfrequenz von 500 im Analysekorpus haben, können 43 Wörter von 130 im WissPubl als Termini identifiziert werden. Wird das gleiche Verfahren auf TranskriptVorl angewandt, aber aufgrund der Korpusgröße mit einer Mindestfrequenz von 100, können fünf Begriffe als Termini angesehen werden. Die Tabelle 5-11 zeigt eine Gegenüberstellung dieser Analyseergebnisse.

	Korpus *WissPubl*	**Korpus *TranskriptVorl***
Gesamtanzahl Tokens	920.237	77.590
Gesamtanzahl Wortformen	8307	1504
Anzahl Wortformen in der 1. Wortformklasse (Klasse der Fachtermini)	5020	250
Anteil vom Korpus	60,43%	16,62%
Anzahl Wortformen in der 2. Wortformklasse	2943	1240
Anteil vom Korpus	35,42%	82,44%
Anzahl Wortformen mit Häufigkeitsquotienten >10	130	57
Anteil von der 2. Wortformklasse	4,42%	4,60%
Anzahl Wortformen mit Häufigkeitsquotienten >10 und Wortfrequenz > 500	43	
Anzahl Wortformen mit Häufigkeitsquotienten >10 und Wortfrequenz > 100		5

Tab. 5-11: Analyse der zweiten Wortformklasse mit Häufigkeitsquotienten

In der Tabelle 5-12 werden einige Begriffe aus jedem Korpus mit dem Häufigkeitsquotienten und der Frequenz im Analysekorpus gezeigt. Die Auflistung aller Wörter, die genannte Kriterien erfüllen, ist in den Anlagen A11 und A12.

	Korpus WissPubl			Korpus TranskriptVorl	
Terminus	Wortfrequenz	Häufigkeits-quotient	Terminus	Wortfrequenz	Häufigkeits-quotient
robot	3495	283,757	potential	167	30,96
sensor	1111	105,410	charge	168	23,01
solar	2011	81,131	structure	173	22,99
efficiency	1571	72,618	power	277	11,36
battery	1405	70,157			
density	618	50,034			
exploration	663	44,449			
energy	5573	42,489	energy	279	25,23
cell	3722	40,599			
load	883	40,219			

Tab. 5-12: Signifikante Termini aus der zweiten Wortformenklasse mit Häufigkeitsquotienten

Der Tabelle 5-12 kann entnommen werden, dass der Terminus *energy* in beiden Korpora enthalten ist. Es folgte eine Überprüfung mittels Differenzanalyse, ob innerhalb der zweiten Wortformenklasse der beiden Korpora gemeinsame Begriffe vorkommen, deren Häufigkeitsquotient höher als 10 liegt. Es wurden 26 gemeinsame Termini gefunden. Einige Beispiele und deren Häufigkeitsquotienten sind: energy (42,48), potential (13), battery (70,15), solar (81,13), efficiency (72,61). Alle 26 gemeinsamen Termini sind in der Anlage A10 aufgelistet.

5.4 Zusammenfassung und Schlussfolgerungen

Die natürliche Sprache ist gekennzeichnet durch Komplexität, Vagheit und Mehrdeutigkeiten. Die Termini eines Wissensgebiets sind Elemente der Fachsprache, bieten einen Zugang zum Wissensgebiet und ermöglichen eine präzise Kommunikation über die fachlichen Konzepte. Text-Mining-Methoden helfen dabei, natürlich sprachliche Texte zu verarbeiten und die relevanten Informationen für einen bestimmten Kontext zu extrahieren und zu verdichten. Pazienza (1998, S. 188) weist auf die Bedeutung der Festlegung von Prinzipien für die Charakterisierung von Termini und für die Beschreibung von grammatikalischen Paradigmen einer Fachsprache bei der Erkennung und Extraktion von Terminologie hin.

Die Analyse der Frequenz von und der statistischen Abhängigkeiten zwischen Wortformen und Wortformenkombinationen unter Berücksichtigung von syntagmatischen und paradigmatischen Relationen sind statistische Methoden, die sich für die Identifizierung von fachspezifischen Termini eignen. Durch die Anwendung von

linguistischen Methoden werden syntaktische und morphologische Muster analysiert, um Termini zu identifizieren. Die hybriden Methoden verbinden statistische und linguistische Verfahren und nutzen deren Stärken, um bestmögliche Ergebnisse bei der Identifizierung und Extraktion von Terminologie zu erzielen.

In der vorliegenden Arbeit wurde die Differenzanalyse für die Terminologie-Extraktion dazu verwendet, um die Häufigkeit der Fachtermini in wissenschaftlicher Fachliteratur und Vorlesungsvideos zu ermitteln und zu vergleichen. Die Ergebnisse der Analyse zeigten, dass im Korpus mit Vorlesungstranskriptionen die Allgemeinsprache einen dominierenden Anteil von 82,44% hat. Über 70% der Fachausdrücke aus diesem Korpus waren im Korpus mit wissenschaftlicher Literatur enthalten. Daraus lässt sich ableiten, dass sich die wissenschaftliche Fachliteratur für das Training des Sprachmodells und die Erweiterung des themenbezogenen Vokabulars der automatischen Spracherkennungssoftware besonders eignet.

Die Analyse der Wortformen, die der Klasse der Termini zugeordnet wurden, hat gezeigt, dass je höher die Frequenz eines Wortes ist, desto höher die Fachspezifik des Wortes ist. Bei den Wortformen mit geringer Frequenz handelt es sich in der Regel um abweichende Schreibweisen oder Rechtschreibfehler. Es wurden jedoch auch in der Mindestfrequenzklasse 2 Fachtermini identifiziert. Durch den Einsatz des Häufigkeitsquotienten konnten weitere Wortformen im Korpus mit Allgemeinwortschatz identifiziert werden, die als Termini angesehen werden können. Wie auch im Kapitel 2.1 gezeigt wurde, werden bei der Bildung von Fachtermini vorhandene Sprachressourcen verwendet, um komplexe Konzepte zu bezeichnen. Die Verwendung von Fachterminologie ist somit kontextabhängig.

Die Ergebnisse zeigen, dass die überwiegende Menge von Termini Nomina sind. Auffallend war die häufige Verwendung von Adjektiven, insbesondere des Adjektivs *high*, mit den Formen *higher* und *highest*. Daraus lässt sich ableiten, dass die meisten fachspezifischen Termini Komposita sind und nach dem Muster Adjektiv–Nomen und Adjektiv–Adjektiv–Nomen gebildet werden. Um solche Wortformen zu identifizieren, sollte die Differenzanalyse auf extrahierte Nominal-

phrasen und nicht allein auf Einwortformen angewandt werden. Empfehlenswert ist hierfür die Anwendung einer hybriden Methode der Textanalyse.

Bei der Verwendung der Differenzanalyse ist die Auswahl des Referenzkorpus besonders wichtig. Die Größe und Repräsentativität der ausgewerteten Publikationen im Referenzkorpus sind zwei Aspekte, die beachtet werden müssen. Die verwendete Liste mit Wortformen aus COCA war mit 5000 Wörtern erheblich zu klein. In so einem Fall kann bei der Klassifizierung von extrahierten Begriffen nicht allein der Grundsatz gelten: „Wörter, die im Referenzkorpus nicht enthalten sind, sind meistens Fachtermini, wenn sie mit einer gewissen Mindestfrequenz auftreten“ (Witschel, 2004, S. 62). Eine Analyse durch Fachexperten oder zusätzliche Kriterien müssen eingesetzt werden, um präzise Ergebnisse zu erzielen. Abhängig vom Ziel der Untersuchung kann die Anschaffung eines umfangreicheren Referenzkorpus oder die selbständige Erstellung einer Wortformenliste mit Frequenzen auf Basis von frei erhältlichen Korpora erfolgen.

In beiden Analysekorpora wurden genannte Entitäten (*Named Entities*) im etwa gleichen Verhältnis identifiziert. Diese Begriffe sind keine Terminologie und daher für die Zwecke dieser Untersuchung nicht relevant. Diese spezifischen Namen könnten schon vorab durch entsprechende POS-Tags (*NE* für Named Entity) bei der Merkmalsanalyse gefiltert und getrennt angezeigt werden (vgl. Witschel, 2004, S. 64). Die große Menge von in Korpora vorhandenen Eigennamen ist dadurch zu erklären, dass in wissenschaftlichen Publikationen, zusätzlich zum Literaturverzeichnis, im Text auf andere relevante Schriften und deren Verfasser hingewiesen wird. Weiterhin kann ein Grund für das Vorhandensein von genannten Entitäten in den nicht vollständig bereinigten Metadaten gesehen werden, welche bei der Konvertierung der PDF-Dokumente mit übernommen wurden.

Hinsichtlich der Reduzierung der Out-of-Vocabulary-Rate und der Erweiterung des Vokabulars der Spracherkennung können die genannten Entitäten durchaus nützlich sein. Die Ergebnisse einer umfassenden Untersuchung von Schaaf (2004, S. 55–60) haben ergeben, dass auch bei Spracherkennern mit großen Vokabularen die genannten Entitäten (und hier besonders die Personennamen) einen wesentlichen Anteil der OOV-Wörter ausmachen und deshalb beachtet werden sollten.

6. Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurde untersucht, ob und wie der prozentuale Anteil der korrekt erkannten Wörter durch Training des Sprachmodells der automatischen Spracherkennung mit fachspezifischer Terminologie wirksam gesteigert werden kann. Hierfür wurde ein Verfahren entwickelt, das die gegebenen Rahmenbedingungen sowie die Anforderungen des Systems für die Durchführung der Domänen-Adaption berücksichtigte. Es konnte gezeigt werden, dass das Web geeignete Ressourcen für das Training eines Sprachmodells bereithält, die zur Verbesserung automatischer Spracherkennung genutzt werden können. Es wurde dargelegt, welche Anzahl und welche Art von Daten benötigt werden, um den Prozentsatz der falsch erkannten Wörter zu senken.

Die Untersuchungsergebnisse zum Domänen-Training haben gezeigt, dass nicht die Menge der Daten für das Training eines Sprachmodells entscheidend ist, sondern die Qualität. Die Adaption eines Sprachmodells kann als kreisläufiger Prozess mit alternierenden Trainings- und Testphasen gestaltet werden, während dessen die Ergebnisse bewertet und Trainingsmethoden angepasst werden. Bei der Evaluierung der Adaption wurden drei Evaluierungsmaße eingesetzt: Perplexität, Wortfehlerrate und Out-of-Vocabulary-Rate. Die Werte dieser Maße geben Hinweise auf die Qualität der Ergebnisse der automatischen Spracherkennung. Die Ergebnisse der Untersuchung haben gezeigt, dass durch Training die Perplexitätswerte deutlich verringert werden können. Im Gegensatz dazu zeigte das Training eine Steigerung der Out-of-Vocabulary-Rate beim Sprachmodell „Informatik“ und nur eine kleine Verringerung beim Sprachmodell „Technik“. Fehlende Wörter im Vokabular des Spracherkenners sind einer der Gründe für Fehler in der Transkription von Sprachaufzeichnungen.

Ausgehend von dieser Feststellung wurde im zweiten Teil der Arbeit untersucht, wieviel fachspezifisches Vokabular in Vorlesungsvideos im Vergleich zu wissenschaftlichen Publikationen benutzt wird, um damit Unterschiede hinsichtlich der verwendeten Fachsprache aufzuzeigen. Um die Häufigkeit der Fachtermini in wissenschaftlicher Fachliteratur und Vorlesungsvideos zu ermitteln und zu vergleichen, wurde die Differenzanalyse verwendet. Die Ergebnisse der Analyse zeigten,

dass im Korpus mit Vorlesungstranskriptionen die Allgemeinsprache einen dominierenden Anteil hat. Über 70% der Fachausdrücke aus dem Korpus mit Vorlesungstranskriptionen waren im Korpus mit wissenschaftlicher Literatur enthalten. Es konnte dadurch gezeigt werden, dass sich die wissenschaftliche Fachliteratur für das Training des Sprachmodells und die Erweiterung des Vokabulars der automatischen Spracherkennung mit Terminologie eignet.

Die Analyse des verwendeten Vokabulars in den Korpora mit Vorlesungsvideos und wissenschaftlichen Publikationen hat gezeigt, dass in beiden Korpora genannte Entitäten (*Named Entities*) im etwa gleichen Verhältnis vorhanden sind.

6.1 Ausblick

Die weiteren Arbeiten an der Adaption des Sprachmodells könnten zukünftig in einigen Punkten effizienter gestaltet werden. Hinsichtlich der Recherche nach relevantem Trainingsmaterial sollten Methoden des Information Retrieval genutzt werden. Mit Hilfe von TF-IDF können relevante Keywords aus den Gebieten identifiziert werden, für die das Sprachmodell trainiert werden sollte. Auf diese Weise ermittelte Suchtermini können bei der Formulierung der Suchanfragen verwendet werden. Fügen (2009, S. 70–72) sowie Lecorvé, Gravier und Sébillot (2008, S. 593–595) zeigen, wie mit dieser Methode aus dem vorhandenen Material geeignete Termini identifiziert und für die Suche verwendet werden können.

Die Extraktion von Suchtermini sollte sich nicht nur auf Einworttermini beschränken. Unter Berücksichtigung der Besonderheiten der technischen Terminologie, die im Kapitel 2 dargestellt wurden, und der im Kapitel 5 dargestellten Ergebnisse der Terminologie-Extraktion, sollte die Identifizierung und Extraktion von Mehrworttermini bevorzugt werden. Diese können dann als Suchtermini eingesetzt werden, um relevantes Trainingsmaterial im Web oder anderen Textkorpora zu finden. Hierfür eignen sich statistische Methoden der Analyse von Kollokationen und Kookkurrenzen, die im Kapitel 5.1.1 beschrieben wurden. Wartena, Brussee und Slakhorst (2010, S. 54–58) zeigen, dass die Extraktion von Schlüsselwörtern

mittels Analyse von Kookkurrenzen durchaus die Ergebnisse von TF-IDF hinsichtlich der Genauigkeit verbessern können.

Die gezielte Auswahl von Schlüsselwörtern, die als Suchtermini eingesetzt werden, kann auch das Filtern und die Auswahl des Trainingsmaterials, das aus den Webressourcen gesammelt wird, effizienter gestalten. Die manuelle Prüfung auf die Eignung jedes einzelnen Dokuments könnte entfallen, denn je präziser die Suchanfrage, desto präziser die Ergebnisse. Unter Berücksichtigung des Ranking-Verfahrens von Websuchmaschinen, kann davon ausgegangen werden, dass je höher eine gefundene Webseite in der Ranking-Liste der Suchmaschine steht, desto relevanter ist sie in Bezug auf die Suchanfrage. Hochqualitative relevante Webseiten können effizient gesucht werden, indem z.B. ein Webspider zum Durchsuchen von ausgewählten Servern eingesetzt wird. Umfangreiche Möglichkeiten in diesem Sinne bietet Lucene[72].

Fehlende Wörter im Vokabular des Spracherkenners rufen Fehler bei der Transkription hervor. Weitere Untersuchungen zur Verbesserung der Transkriptionsergebnisse könnten darin bestehen, die Out-of-Vocabulary-Wörter nach der Adaption zu untersuchen und deren Aufnahme im Wörterbuch des Spracherkenners zu ermöglichen. Wie die Ergebnisse der Terminologie-Extraktion gezeigt haben, ist der Anteil von Named Entities sowohl in den Fachpublikationen als auch in den Vorlesungsvideos beachtlich.

[72] http://lucene.apache.org/

Literaturverzeichnis

Ahmad, Khurshid. (1997). Corpus Linguistics and Terminology Extraction. *Application-oriented Terminology Management*, Handbook of terminology management (Vol. 2, pp. 419–425). Amsterdam; Philadelphia: J. Benjamins.

Ahmad, Khurshid, & Rogers, Margaret. (1997). Corpus Linguistics and Terminology Extraction. *Application-oriented Terminology Management*, Handbook of terminology management (Vol. 2, pp. 725–760). Amsterdam; Philadelphia: J. Benjamins.

Armstrong, Christopher J., & Large, Andrew. (2001). Search strategies: some general considerations. In Armstrong, Christopher J. & Large, Andrew (Eds.), *Manual of Online Search Strategies* (3. ed., Vol. 1: Sciences, pp. 1–22). Aldershot: Gower.

Arppe, Antti. (1995). Term extraction from unrestricted text : NODALIDA-95. Elektronische Ressource verfügbar unter URL: www2.lingsoft.fi/doc/nptool/term-extraction.html [Zuletzt geprüft am 27.02.13]

Bargheer, Margo, Bellem, Saskia, & Schmidt, Birgit. (2006). Open Access und Institutional Repositories - Rechtliche Ramenbedingungen. In Spindler, Gerald (Ed.), Rechtliche Rahmenbedingungen von Open Access-Publikationen., Göttinger Schriften zur Internetforschung (Vol. 2, pp. 1–20). Göttingen: Universitätsverlag Göttingen.

Bernhard, Delphine. (2006). Multilingual Term Extraction from Domain-specific Corpora Using Morphological Structure. EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, April 3-7, 2006, Trento, Italy (pp. 171–174). Elektronische Ressource verfügbar unter URL: http://acl.ldc.upenn.edu/E/E06/E06-2022.pdf [Zuletzt geprüft am 10.02.13]

Bontcheva, Kalina, Tablan, Valentin, Maynard, Diana, & Cunningham, Hamish. (2004). Evolving GATE to meet new challenges in language engineering. Natural Language Engineering, 10(3-4), 349–373.

Borgman, C. L. (1996). Why are online catalogs still hart to use? Journal of the American Society for Information Science, 47(7), 493–503.

Bulyko, Ivan, Stolcke, Andreas, & Ostendorf, Mari. (2003). Getting more Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures. Proceedings of 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics; May 27 to June 1, 2003, Edmonton, Alberta, Canada; HLT-NAACL 2003, Vol. 2, 7–9. Elektronische Ressource verfügbar unter URL: http://www.aclweb.org/anthology/N/N03/N03-2003.pdf [Zuletzt geprüft am 27.02.13]

Chang, G., Healey, Marcus J., McHugh, James A. M., & Wang, Jason T. L. (2001). Mining the World Wide Web : an information search approach. Boston: Kluwer Academic.

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. System, 32(2), 251–263. Elektronische Ressource verfügbar unter
URL: http://dx.doi.org/10.1016/j.system.2003.11.008 [Zuletzt geprüft am 27.02.13]

Cunningham, H., Bonceva, K., & Maynard, D. (2011). Text Processing with GATE. Sheffield: University of Sheffield Dept. of Computer Science. [Computer Software]. Elektronische Ressource verfügbar unter URL: http://gate.ac.uk/gate/doc/papers.html [Zuletzt geprüft am 27.02.13]

Datenbank Infosystem (DBIS). Homepage. Elektronische Ressource verfügbar unter URL: http://rzblx10.uni-regensburg.de/dbinfo/fachliste.php?lett=l [Zuletzt geprüft am 27.02.13]

Davies, Mark. (n.d.). Corpus of Contemporary American English. Elektronische Ressource verfügbar unter URL: http://www.wordfrequency.info/top5000.asp [Zuletzt geprüft am 27.02.13]

Dubuc, Robert, & Lauriston, Andy. (1997). Terms and Contexts. Basic aspects of terminology management, Handbook of terminology management (Vol. 1, pp. 80–87). Amsterdam; Philadelphia: J. Benjamins.

Felber, H., & Budin, G. (1989). Terminologie in Theorie und Praxis. Forum für Fachsprachen-Forschung (Vol. 9). Tübingen: Narr.

Frantzi, K. T., & Ananiadou, S. (1996). Extracting nested collocations. *Proceedings of COLING* (pp. 41–46). Copenhagen. Elektronische Ressource verfügbar unter URL: http://acl.ldc.upenn.edu/C/C96/C96-1009.pdf [Zuletzt geprüft am 20.02.13]

Fügen, C. (2009). A System for Simultaneous Translation of Lectures and Speeches. Karlsruhe Institute of Technology, Karlsruhe. Elektronische Ressource verfügbar unter URL: http://digbib.ubka.uni-karlsruhe.de/volltexte/1000013594 [Zuletzt geprüft am 27.02.13]

Gildea, Daniel, & Hofmann, Thomas. (1999). Topic-based language models using EM. Proceedings of the Eurospeech '99, 6th European Conference on Speech Communication and Technology, Budapest, Hungary, 5 - 9 September 1999, European Speech Communication Association (ESCA) (pp. 2167–2170). Bonn: ESCA.

Glass, J., Barzilay, R., Huynh, D., Hazen, T. J., Cyphers, S., & Malioutov, I. (2007). Recent Progress in the MIT Spoken Lecture Processing Project. 8th annual conference of the International Speech Communication Association : Interspeech 2007; August 27 - 31, 2007, Antwerp, Belgium (Vol. Vol. 3, pp. 2392–2395). Red Hook, NY: Curran.

Griesbaum, Joachim, Bekavac, Bernard, & Rittberger, Marc. (2009). Typologie der Suchdienste im Internet. In Lewandowski, Dirk (Ed.), Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis (pp. 18–52). Heidelberg: AKA, Akad. Verl.-Ges.

Granitzer, Michael. (2006). Statistische Verfahren der Textanalyse. In Pellegrini, Tassilo & Blumauer, Andreas (Eds.), Semantic Web: Wege zur vernetzten Wissensgesellschaft (pp. 437–451). Berlin, Heidelberg: Springer.

Handreck, Franka, & Mönnich, Michael W. (2008). Google Scholar als Alternative zu wissenschaftlichen Fachdatenbanken. *B.I.T. online*, *4*, 401–412. Elektronische Ressource verfügbar unter URL: http://shan01.tib.uni-hannover.de/han/2384/www.b-i-t-online.de/archiv/2008-04-idx.html [Zuletzt geprüft am 27.02.13]

Hauptmann, Alexander, Jin, Rong & Wactlar, Howard. (2000). Data Analysis for a Multimedia Library. (Renals, S. & Grefenstette, G., Eds.) Text and speech triggered information access: 8th ELSNET Sommer School, Chios Island, Greece, Juli 15-30 2000, Lecture Notes in Artificial Intelligence, 2705, 6–37.

Heid, Ulrich. (1997). Collocations in Sublanguage Texts: Extraction from Corpora. Application-oriented Terminology Management, Handbook of terminology management (Vol. 2, pp. 788–808). Amsterdam; Philadelphia: J. Benjamins.

Heyer, G., Quasthoff, U., & Wittig, T. (2006). Text Mining: Wissensrohstoff Text : Konzepte, Algorithmen, Ergebnisse. Herdecke [u.a.]: W3L-Verl.

Hoffmann, L. (1988). Vom Fachwort zum Fachtext: Beiträge zur Angewandten Linguistik. S.l.: G. Narr.

Jurafsky, D., & Martin, J. (2009). Speech and Language Processing (Second ed.). Upper Saddle River, New Jersey: Pearson Education.

Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(01), 9–27. Elektronische Ressource verfügbar unter URL: http://dx.doi.org/10.1017/S1351324900000048 [Zuletzt geprüft am 27.02.13]

Kind, Joachim. (2004). Praxis des Information Retrievals. In Kuhlen, Rainer (Ed.), Handbuch zur Einführung in die Informationswissenschaft und -praxis, Grundlagen der praktischen Information und Dokumentation (5., völlig neu gefasste Ausg., Vol. 1, pp. 389–398). München: Saur.

Kneser, R., & Peters, J. (1997). Semantic clustering for adaptive language modeling. 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, April 21-24, 1997, Munich, Germany; [proceedings] (Vol. 2: Speech processing, pp. 779–782). Los Alamitos, CA, USA: IEEE Comput. Soc. Press. Elektronische Ressource verfügbar unter URL: http://dx.doi.org/10.1109/ICASSP.1997.596041 [Zuletzt geprüft am 27.02.13]

Kompetenzzentrum für nicht-textuelle Materialien an der Technischen Informationsbibliothek Hannover. Homepage. Elektronische Ressource verfügbar unter URL: http://www.tib-hannover.de/de/dienstleistungen/kompetenzzentrum-fuer-nicht-textuelle-materialien-knm/ [Zuletzt geprüft am 27.02.13]

Konchady, M. (2008). Building search applications : Lucene, LingPipe, and Gate. Oakton, VA: Mustru Pub.

Lamel, L., Bilinski, E., Adda, G., Schwenk, H., Gauvain, J.-L., & Steve Renals, Samy Bengio, and Jonathan Fiskus, editors. (2006). The LIMSI RT06s Lecture Transcription System. Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA, Lecture Notes in Computer Science (Vol. 4299, pp. 457–468). Berlin/Heidelberg: Springer Verlag.

Lecorvé, Gwénolé, Gravier, Guillaume, & Sébillot, Pascale. (2008). On the Use of Web Resources and Natural Language Processing Techniques to Improve Automatic Speech Recognition Systems. LREC 2008 (pp. 592–599). Presented at the Language Resources and Evaluation Conference; LREC ; 6 (Marrākuš) : 2008. Elektronische Ressource verfügbar unter URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/155_paper.pdf [Zuletzt geprüft am 10.02.13]

Lewandowski, Dirk. (2009). Spezialsuchmaschinen. In Lewandowski, Dirk (Ed.), Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis (pp. 53–70). Heidelberg: AKA, Akad. Verl.-Ges.

Mahajan, M., Beeferman, D., & Huang, X. D. (1999). Improved topic-dependent language modeling using information retrieval techniques. Proceedings / 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing: ICASSP 99; March 15 - 19, 1999, Civic Plaza, Hyatt Regency, Phoenix, Arizona, U.S.A. (Vol. Vol.1: Speech processing 1, pp. 541–544). Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing ; 1999 (Phoenix, Ariz.) : 1999.03.15-19, Piscataway, NJ: IEEE Service Center. Elektronische Ressource verfügbar unter URL: http://dx.doi.org/10.1109/ICASSP.1999.758182 [Zuletzt geprüft am 12.02.13]

Manning, C. D., & Schütze, H. (2001). Foundations of statistical natural language processing. Cambridge, Mass.: MIT Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. New York: Cambridge University Press.

McKeating, Stephanie, & MacLeod, Roddy. (2001). Engineering and energy. In Armstrong, Chris & Large, Andrew (Eds.), Manual of Online Search Strategies (3. ed., Vol. 1: Sciences, pp. 279–329). Gower.

Meinel, Christoph, & Sack, Harald. (2009). Prolog. Digitale Kommunikation: Vernetzen, Multimedia, Sicherheit, X.Media.Press (pp. 1–17). Berlin: Springer-Verlag.

Munteanu, Cosmin. (2009). Useful Transcriptions of Webcast Lectures (Diss.). University of Toronto, Department of Computer Science, Toronto. Elektronische Ressource verfügbar unter URL: http://www.cs.toronto.edu/~mcosmin/publications/phdthesis.pdf [Zuletzt geprüft am 27.02.13]

LM Workplace: Users Guide. (2012). Version 1.0. [Unveröffentlichtes Dokument]
Pazera, Jacek. (2010). Pazera Free Audio Extraktor. [Computer Software]. Elektronische Ressource verfügbar unter URL: http://www.pazera-software.com/products/audio-extractor/ [Zuletzt geprüft am 27.02.13]

Pazienza, M. T. (1998). A domain-specific terminology-extraction system. Terminology, 5(2), 183–201. Elektronische Ressource verfügbar unter
URL: http://dx.doi.org/:10.1075/term.5.2.07paz [Zuletzt geprüft am 27.02.13]

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology Extraction: an Analysis of Linguistic and Statistical Approaches. In S. Sirmakessis (Ed.), Knowledge Mining, Studies in Fuzziness and Soft Computing (Vol. 185, pp. 255–276). Berlin/Heidelberg: Springer-Verlag. Elektronische Ressource verfügbar unter URL: http://www.springerlink.com/index/10.1007/3-540-32394-5_20 [Zuletzt geprüft am 27.02.13]

Pieper, Dirk, & Wolf, Sebastian. (2009). Wissenschaftliche Dokumente in Suchmaschinen. In Lewandowski, Dirk (Ed.), Handbuch Internet-Suchmaschinen: Nutzerorientierung in Wissenschaft und Praxis (pp. 356–374). Heidelberg: AKA, Akad. Verl.-Ges.

Poetzsch, E. (2006). Information Retrieval: Einführung in Grundlagen und Methoden. Berlin: Poetzsch.

Rösch, Hermann, & Weisbrod, Dirk. (2004). Linklisten, Subject Gateways, Virtuelle Fachbibliotheken, Bibliotheks- und Wissenschaftsportale: Typologischer Überblick und Definitionsvorschlag. B.I.T. *Online*, *3*, 177–188. Elektronische Ressource verfügbar unter URL: http://shan01.tib.uni-hannover.de/han/2384/www.b-i-t-online.de/archiv/2004-03-idx.html [Zuletzt geprüft am 27.02.13]

Rothkegel, A. (2009). Technikkommunikation. UTB (Vol. 3214: Sprachwissenschaft). Stuttgart: UTB.

Sager, J. C., Dungworth, D., & McDonald, P. F. (1980). English special languages : principles and practice in science and technology. Wiesbaden: Brandstetter.

Sager, Juan C. (1997). Term Formation. Basic aspects of terminology management, Handbook of terminology management (Vol. 1, pp. 25–41). Amsterdam; Philadelphia: J. Benjamins.

Schaaf, Thomas. (2004). Erkennen und Lernen neuer Wörter. Univ. Karlsruhe, Diss., Karlsruhe. Elektronische Ressource verfügbar unter URL http://digbib.ubka.uni-karlsruhe.de/volltexte/1000001684 [Zuletzt geprüft am 27.02.13]

Schmalenbach, K. (2000). Terminlologie - Grundbegriffe. Fachzeitschrift Technische Dokumentation, (11). Elektronische Ressource verfügbar unter
URL: http://www.doku.net/techndoku/artikel/terminolo0.htm [Zuletzt geprüft am 27.02.13]

Sethy, Abhinav, Narayanan, Shrikanth, & Georgiou, Panayiotis G. (2005). Building Topic Specific Language Models from Webdata using Competitive Models. Proceedings of the European Conference on Speech Communication and Technology (INTERSPEECH), Lisbon, Portugal, 2005. ISCA., 1293–1296.

Stouten, F., Duchateau, J., Martens, Jean-Pierre, & Wambacq, Patrick. (2006). Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. Speech Communication, 48, 1590–1606.

Sühl-Strohmenger, Wilfried. (2008). Digitale Welt und Wissenschaftliche Bibliothek - Informationspraxis im Wandel. Bibliotheksarbeit (Vol. 11). Wiesbaden: Harrassowitz.

Suzuki, Hisami, & Gao, Jianfeng. (2005). A Comparative Study on Language Model Adaptation Techniques Using New Evaluation Metrics. In Association for Computational Linguistics (Ed.), Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (pp. 265–272). Vancouver, British Columbia, Canada: Association for Computational Linguistics. Elektronische Ressource verfügbar unter
URL: http://www.aclweb.org/anthology/H/H05/H05-1034 [Zuletzt geprüft am 20.02.13]

Wan, V., & Hain, T. (2006). Strategies for Language Model Web-Data Collection. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, 2006. IEEE., 1069–1072.

Wartena, C., Brussee, R., & Slakhorst, W. (2010). Keyword Extraction Using Word Co-occurrence. In A. M. Tjoa (Ed.), Workshops on Database and Expert Systems Applications, 30 August - 3 September 2010, Bilbao, Spain (pp. 54–58). Presented at the International Workshop on Database and Expert Systems Applications ; 21 (Bilbao, Spain) : 2010.08.30-09.03, Los Alamitos, Calif.: IEEE Computer Society. Elektronische Ressource verfügbar unter

URL: http://doi.ieeecomputersociety.org/10.1109/DEXA.2010.32 [Zuletzt geprüft am 20.02.13]

Witschel, H. F. (2004). Terminologie-Extraktion : Möglichkeiten der Kombination statistischer und musterbasierter Verfahren. Content and communication (Vol. 1). Würzburg: Ergon Verlag.

Wright, Sue Ellen. (1997). Term Selection: The Initial Phase of Terminology Management. Basic aspects of terminology management, Handbook of terminology management (Vol. 1, pp. 13–23). Amsterdam; Philadelphia: J. Benjamins.

Wüster, E. (1979). Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Wien: Springer.

Zhu, X., & Rosenfeld, R. (2001). Improving trigram language modeling with the world wide web. Speech processing 1, Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 26 (Salt Lake City, Utah) : 2001.05.07-11, ICASSP 2001 ; 26, 533–536.

Anlagen A

Terminologie-Extraktion

Anlage A1

Dokumentation der verwendeten Vorlesungstranskriptionen zwecks Terminologie-Extraktion

	Institution	Fach	Typ	Vorlesungsreihe	Reihe-Nr.	Vorle-sung	Titel	Sprecher	Datum	Dauer	Tokens	URL
1.	Massachusetts Institute of Technology (MIT)	Physik	Video Lecture	Physics I	8.01 Classical Mechanics	Lecture 14	Orbits and Escape Velocity	Lewin, Walter	13.10.1999 (veröffentlicht am 03.01.2008)	00:50:06	7112	http://ocw.mit.edu/courses/physics/8-01-physics-i-classical-mechanics-fall-1999/video-lectures/lecture-14/
2.	Massachusetts Institute of Technology (MIT)	Chemie	Video Lecture	Principles of chemical science II	5.112	Lecture 23	Cell potentials and free energy	Cummins, Christopher	Herbst 2005 (veröffentlicht am 08.05.2007)	00:43:08	5234	http://ocw.mit.edu/courses/chemistry/5-112-principles-of-chemical-science-fall-2005/video-lectures/lecture-23-cell-potentials-and-free-energy/
3.	Massachusetts Institute of Technology (MIT)	Physik	Video Lecture	Physics II	8.02 Electricity & Magnetism	Lecture 7	Capacitance and Field Energy	Lewin, Walter	Frühjahr 2002 (veröffentlicht am 18.05.2007)	00:49:26	7435	http://ocw.mit.edu/courses/physics/8-02-electricity-and-magnetism-spring-2002/video-lectures/lecture-7-capacitance-and-field-energy/
4.	Massachusetts Institute of Technology (MIT)	Physik	Video Lecture	Physics II	8.02 Electricity & Magnetism	Lecture 10	Batteries and EMF	Lewin, Walter	Frühjahr 2002 (veröffentlicht am 18.05.2007)	00:50:05	6715	http://ocw.mit.edu/courses/physics/8-02-electricity-and-magnetism-spring-2002/video-lectures/lecture-10-batteries-and-emf/
5.	Massachusetts Institute of Technology (MIT)	Aeronautik & Astronautik	Video Lecture	Aircraft Systems Engineering	16.885 J	Lecture 21	Systems Engineering for Space Shuttle Payloads	Lavoie, Anthony	Herbst 2005 (veröffentlicht am 09.06.2009)	01:52:07	15430	http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-885j-aircraft-systems-engineering-fall-2005/video-lectures/lecture-21/#vid_related
6.	Massachusetts Institute of Technology (MIT)	Electrotechnik	Video Lecture	Electrical Engineering and Computer Science	6.002 Circuits and Electronics	Lecture 22	Energy and power	Agarwal, Anant	Frühjahr 2007 (veröffentlicht am 18.05.2007)	00:51:41	6593	http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-002-circuits-and-electronics-spring-2007/video-lectures/6_0022007L022.pdf
7.	Massachusetts Institute of Technology (MIT)	Electrotechnik	Video Lecture	Electrical Engineering and Computer Science	6.002 Circuits and Electronics	Lecture 23	Energy, CMOS	Agarwal, Anant	Frühjahr 2007 (veröffentlicht am 18.05.2007)	00:40:11	4807	http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-002-circuits-and-electronics-spring-2007/video-lectures/6_0022007L023.pdf
8.	The British Academic Spoken English Corpus (BASE)	Engineering	Vorlesung/Seminar	Modul: Heat transfer	k.A.	k.A.	Renewable energy	24 Personen (Keine Angabe von Namen	08.03.2000	00:45:17	5546	http://www.reading.ac.uk/AcaDepts/ll/base_corpus/ps/index.htm
9.	The British Academic Spoken English Corpus (BASE)	Engineering	Vorlesung/Seminar	Modul: unbekannt	k.A.	k.A.	Tension structures	2 Personen (Keine Angabe von Namen)	01.12.1998	00:54:21	6939	http://www.reading.ac.uk/AcaDepts/ll/base_corpus/ps/index.htm
10.	Australian National University	Engineering	Vorlesung	Public Lecture Series	k.A.	k.A.	Solar thermal concentrators: Capturing the sun for large scale power generation and energy export	Lovegrove, Keith	07.04.2008	k.A.	11779	http://www.science.org.au/events/publiclectures/re/lovegrove.html
											77590	

Massachusetts Institute of Technology (MIT): Transkriptionen veröffentlicht unter der Lizenz CC BY-NC-SA 3.0 U.S.

BASE Transkriptionen sind frei verfügbar für wissenschaftliche Zwecke unter folgenden Bedingungen: keine vollständige Wiedergabe für großes Publikum weder innerhalb von Publikationen noch während Lehrveranstaltungen; für Forschungszwecke ist die Nutzung von kleinen Teilen erlaubt; Reproduktionen sind untersagt; die Entwickler sind zu informieren über jede Publikation, die die Analyse von BASE-Corpus erwähnt.
(Übersetzung der Verfasserin; Quelle: http://www.reading.ac.uk/AcaDepts/ll/base_corpus/base_manual.pdf)

Anlage A 2

Übersicht mit den 25 häufigsten Wörtern aus den Analysekorpora

	WissPubl 1. Klasse der Fachtermini			***TranskriptVorl*** 1. Klasse der Fachtermini		
Rang	Terminus	Dokument Frequenz	Wort Frequenz	Terminus	Dokument Frequenz	Wort Frequenz
1	electrode	36	1421	capacitor	3	116
2	*datum*	127	1401	*voltage*	8	88
3	lithium	35	854	volt	5	79
4	simulation	64	819	zero	8	78
5	duct	18	721	minus	8	59
6	*thermal*	93	701	watt	4	58
7	parameter	74	696	*thermal*	3	52
8	renewable	58	656	joule	3	50
9	*voltage*	51	603	resistor	3	44
10	electrolyte	29	598	capacitance	3	41
11	*discharge*	32	593	inverter	2	40
12	*higher*	99	575	*discharge*	5	37
13	silicon	16	573	dissipate	4	37
14	configuration	53	560	metre	3	36
15	ventilation	30	533	*higher*	8	29
16	interface	53	522	ion	2	28
17	electrochemical	30	487	percent	3	28
18	leakage	17	460	multiply	7	27
19	savings	56	457	radius	4	27
20	controller	40	452	ampere	2	26
21	algorithm	46	413	thickness	2	26
22	robotic	28	409	*datum*	3	25
23	further	99	403	copper	3	24
24	heating	70	388	uranium	3	23
25	plenum	9	386	kilometer	3	21

Anlage A 3

Übersicht mit den extrahierten Wörtern aus der Klasse der potentiellen Fachtermini, die als nicht fachspezifisch eingestuft wurden

Korpus WissPubl

Kategorien	Anzahl	Beispiele
Geografische Namen	122	Minnesota, Belgium, Barcelona, Pennsylvania, Ohio, Frankfurt, Kentucky, Shanghai, Alaska, Munich, Quebec, Portugal, Stockholm, Honolulu, Yale, San Francisco, Boston, Zealand, Taiwan, Arlington, Carolina, Arizona
Personennamen	140	Reagan, Obama, Bruemmer, Verbrugge, Douglas, Piette, Wang, Hamilton, Richards, Zimmerman, Braun, Wong, Andrew, Bryan, Agrawal, Jefferson, Gupta, Jenkins, Kennedy, Planck, Aaron, Clinton, Reagan
Monatsnamen	10	April, August, October, Juli, December
Körperschaftsnamen	17	Wiley, Elsevier, Harvard, Princeton, Stanford, Intel, Carnegie, Sony, Microsoft
Bezeichnung von Software/Datenbanken	9	ScienceDirect, SciVerse, Linux, Google, Hewlett-Packard,
Fremdsprachige Wortformen	8	GmbH, Verlag, Siegel, Universita, Universidad, ecucaccio, instituto,
Metadaten	10	https, copyright, handbook, PDF, dd-mm-yyyy, mailto, issn
Adjektive	20	higher, greater, larger, shorter, easier, worse, worst, bigger, biggest, latest, longest, harder, earliest
Zahlen	8	21th, 10th, 28th, ninety
	344	

Korpus TranskriptVorl

Kategorie	Anzahl	Beispiele
Geografische Namen	5	Europe, California, Japan, Canberra, Germany
Personennamen	3	Geiger, Lewin, Peter
Monatsnamen	2	October, February
Wochentage	3	Tuesday, Thursday, Monday
Währung	1	dollar
	14	

Anlage A4 Extrahierte nicht fachspezifische Wortformen aus der Klasse der Fachtermini im Korpus WissPubl

Wortform	Dok. Frequenz	Wort-frequenz	Wortform	Dok. Frequenz	Wort-frequenz	Wortform	Dok. Frequenz	Wort-frequenz
further	99	403	nearest	8	21	australian	4	10
california	38	301	george	5	21	bologna	2	10
greater	58	200	italy	6	21	francisco	5	9
larger	55	195	mary	10	21	yale	2	9
simon	5	158	kyoto	5	20	honolulu	3	9
john	27	81	kurtz	8	19	peter	6	9
copyright	19	78	bigger	9	19	chris	5	9
december	22	67	worst	15	18	huang	7	8
herein	28	65	reagan	2	18	england	4	8
walker	13	61	firstly	12	18	taiwan	8	8
october	16	57	switzerland	9	17	arlington	2	8
lawrence	22	56	piette	8	17	latest	8	8
washington	13	55	georgia	5	17	stockholm	3	8
august	21	54	richard	8	17	william	8	8
canada	19	54	pacific	10	17	beijing	6	8
wang	11	53	douglas	6	16	timothy	2	7
november	21	52	johnson	5	16	columbia	5	7
gmbh	9	52	india	6	16	jessica	4	7
wiley	3	48	robert	8	16	calif	2	7
elsevier	19	47	massachusetts	9	15	pasadena	6	7
largest	26	47	france	8	15	toulouse	2	7
texas	13	46	wileyonlinelibrary	3	15	andrews	2	7
hewlett	2	46	google	5	14	nielsen	2	7
verlag	5	44	sarah	9	14	russia	5	7
david	19	44	zhang	7	14	carolina	5	7
china	16	42	james	9	14	chicago	3	7
july	21	41	siegel	4	14	karin	2	7
america	15	39	clinton	2	14	arthur	3	6
idaho	7	39	stanford	7	13	eric	5	6
weinheim	3	38	martin	9	13	mathieu	2	6
colorado	11	38	biggest	8	13	luis	2	6
wiley-vch	3	38	korea	4	13	wiley-	3	6
april	16	37	scott	11	13	microsoft	5	6
orient	14	36	thomas	8	13	denver	5	6
shorter	25	36	ontario	6	12	nordman	3	6
florida	10	36	harvard	9	12	quebec	2	6
february	13	35	princeton	3	12	sony	4	6
easier	25	34	dickerhoff	6	12	russell	3	6
september	21	34	kramer	6	12	portugal	4	6
germany	14	33	mexico	7	11	jack	5	6
january	21	31	aaron	8	11	phillip	2	6
austin	6	31	yorktown	2	11	madrid	3	6
york	11	30	asia	5	11	harris	4	6
spain	10	29	wilson	5	11	craig	2	6
saudi	4	29	obama	2	11	phillips	4	6
bruemmer	5	28	jerry	8	11	ottawa	2	6
houston	5	26	handbook	5	10	nevada	6	6
verbrugge	5	25	denmark	4	10	boston	5	5
japan	11	25	jersey	3	10	brazil	3	5
vegas	4	24	sweden	6	10	charlene	5	5
friedman	11	24	steven	3	10	tokyo	4	5
issn	4	23	alan	8	10	elena	4	5
sacramento	3	23	paul	8	10	robinson	2	5
montreal	4	23	jeff	9	10	vienna	4	5
worse	20	23	bruce	7	10	marie	2	5
arabia	4	23	sciencedirect	10	10	michelle	5	5
daniel	17	22	africa	6	10	sunday	2	5
orlando	13	22	cambridge	5	10	longest	5	5

Anlage 4 - Fortsetzung

Wortform	Dokument-Frequenz	Wort-frequenz	Wortform	Dokument-Frequenz	Wort-frequenz	Wortform	Dokument-Frequenz	Wort-frequenz
cheng	4	5	stephen	3	3	baltimore	2	2
mailto	5	5	malibu	3	3	detroit	2	2
linux	2	5	manuel	3	3	hamilton	2	2
hawaii	4	5	pittsburgh	2	3	niagara	2	2
philips	2	5	horowitz	2	3	wilmington	2	2
brisbane	2	5	zahler	3	3	srivastavaetal	2	2
toyota	3	5	helmut	2	3	anderson	2	2
jonathan	4	5	thankful	2	3	kennedy	2	2
roberts	3	5	warburg	3	3	willliam	2	2
aurbach	3	5	tennessee	2	3	kenneth	2	2
wisconsin	4	5	thailand	3	3	keith	2	2
mitsubishi	4	4	ryan	3	3	oklahoma	2	2
verbruggeb	4	4	winkelmann	2	3	reister	2	2
atlanta	3	4	carnegie	2	3	nathan	2	2
alessandro	3	4	iowa	3	3	prijyanonda	2	2
dd-mm-yyyy	4	4	kansas	3	3	copenhagen	2	2
virginia	2	4	britain	3	3	jennie	2	2
chernobyl	2	4	polish	3	3	portuguese	2	2
morris	3	4	swedish	3	3	kreider	2	2
fraunhofer	2	4	miami	2	3	edward	2	2
arizona	4	4	jeffrey	3	3	instituto	2	2
sciverse	4	4	schweiger	2	3	taipei	2	2
northwestern	2	4	smith	3	3	werner	2	2
ohio	3	4	zheng	2	3	15th	2	2
brenner	4	4	seattle	2	3	paris	2	2
michel	3	4	mellon	2	3	alaska	2	2
zealand	2	4	whittingham	2	3	zimmerman	2	2
brian	4	4	universita	2	3	albuquerque	2	2
toronto	3	4	28th	3	3	richards	2	2
pdf	3	4	universidad	2	3	mbraun	2	2
janet	4	4	carlos	3	3	maryland	2	2
norman	2	4	munich	2	3	wong	2	2
harder	3	4	pennsylvania	3	3	southwestern	2	2
co-author	4	4	frankfurt	3	3	karl	2	2
netherlands	3	4	steiner	2	3	24th	2	2
https	2	4	kentucky	3	3	jahn	2	2
illinois	2	4	sixteen	2	3	andrew	2	2
hitachi	3	4	victoria	2	3	bryan	2	2
norford	3	4	shanghai	3	3	planck	2	2
julie	2	4	zhou	2	3	ninety	2	2
williams	3	4	minnesota	3	3	wagner	2	2
harry	4	4	belgium	3	3	lexington	2	2
andy	4	4	gupta	3	3	utah	2	2
donald	2	4	barcelona	2	3	agrawal	2	2
springer	4	4	jenkins	2	3	jefferson	2	2
andreas	3	4	honda	2	3	kingston	2	2
21st	4	4	denis	2	3	oregon	2	2
10th	2	4	monica	2	3	gregory	2	2
riyadh	2	4	fischer	3	3	earliest	2	2
alexander	3	4	jennifer	3	3	twenty-	2	2
south-east	2	4	stuttgart	3	3	matthew	2	2
geneva	3	4	oakland	2	3	earlier	37	107
berlin	3	4	educacio	2	2	longer	45	96
klein	3	4	brunauer	2	2			
louis	2	4	philippines	2	2			
stefan	2	4	srwastava	2	2			
whitney	2	4	croatia	2	2			
intel	3	4	philadelphia	2	2			

Anlage A 5

Extrahierte nicht fachspezifische Wortformen aus der Klasse der Fachtermini im Korpus TranskriptVorl

Wortform	Dokument-Frequenz	Wort-frequenz
lewin	2	14
europe	2	10
california	2	7
japan	3	5
canberra	2	4
dollar	2	4
peter	2	4
february	2	3
germany	2	3
monday	2	3
october	3	3
tuesday	2	3
geiger	2	2
thursday	2	2

Anlage A 6

4839 im WissPubl enthaltene Fachtermini nach Frequenzklassen

Frequenzklassen	Anzahl Wörter	Beispiele von Termini, die in der jeweiligen Gruppe die höchste Frequenz haben
maximal 2	965	Nanoarchitecture, narbonne, trochoidal, isocyanate
zwischen 3 und 10	2341	single-junction, polymerization, nano-structured, volts,
zwischen 11 und 30	976	Nanotechnology, oscillatory, extrapolation, microelectrode, lattice-matching, fault-tolerance, nanometer-sized
zwischen 31 und 50	254	Encoder, depressurization, robots, electrolysis, polynomial, dimethyl, spectroscopy
zwischen 51 und 80	125	Contaminant, li-air, transmitter, multijunction, human-robot, nanotube, humanoid
zwischen 81 und 100	46	Triple, airflow, synthesis, hysteresis, membrane, greenhouse, pre-cooling
zwischen 101 und 150	61	Lithiation, pollutant, cooling, emitter
zwischen 151 und 200	28	Porosity, simulator, semiconductor, conductivity
zwischen 201 und 300	22	Electron, calibration, gaas, subcell
zwischen 301 und 400	9	Cathode, substrate, trajectory
zwischen 401 und 500	5	Electrochemical, leakage, savings, controller, robotic
zwischen 501 und 1000	7	Lithium, simulation, duct, parameter, ventilation
und darüber	1	Electrode (Häufigkeit=1421)

Anlage A 7

Beispiele von 25 Termini jeder Frequenzklasse aus dem Korpus WissPubl

Frequenzklasse: max. 2 Vorkommen			Frequenzklasse: zwischen 3-10		
Terminus	**Dokument-frequenz**	**Wort-frequenz**	**Terminus**	**Dokument-frequenz**	**Wort-frequenz**
field-powered	2	2	notation	8	10
soiling	2	2	glazed	5	10
current-mode	2	2	maximise	5	10
oberlin	2	2	warren	9	10
oscilloscope	2	2	symmetry	5	10
smaller-scale	2	2	summarised	2	10
brittle	2	2	vfmin	2	10
constructing	2	2	simplest	7	10
ramp-up	2	2	cpuc	2	10
reside	2	2	ward	6	10
core-coil	2	2	two-dimensional	4	10
platform-independent	2	2	sandia	5	10
passively	2	2	scaling	5	10
indict	2	2	ptfe	4	10
screenshot	2	2	printing	5	10
gaming	2	2	srivastavaet	2	10
surftest-	2	2	euis	2	10
optimizationstrategy	2	2	scheduling	9	10
emergence	2	2	characterised	3	10
underpredict	2	2	degrade	9	10
elbestawlt	2	2	saft	3	10
requiring	2	2	unbalanced	4	10
radiance	2	2	referenced	9	10
lkvfdus	2	2	puma	4	10
railway	2	2	publications	6	10
peculiar	2	2	cancellation	7	10

Anlage A 7a

Beispiele von 25 Termini jeder Frequenzklasse aus dem Korpus WissPubl

Frequenzklasse: zwischen 11-30			Frequenzklasse: zwischen 11-50		
Terminus	Dokument-frequenz	Wort-frequenz	Terminus	Dokument-frequenz	Wort-frequenz
modelled	5	30	seasonal	8	50
visualization	4	30	setback	6	50
simulations	8	30	quadrotor	2	50
unoccupied	13	30	acute	6	50
sulfur	6	30	spatial	16	50
ceramic	8	30	parameters	12	50
oscillation	13	30	surrogate	3	50
measurements	6	30	reset	10	50
test-bed	2	30	positioning	15	50
usefulness	26	30	gradient	19	50
hierarchical	7	30	closing	8	50
high-speed	6	30	beneficial	20	49
onboard	8	30	modification	28	49
revise	25	30	scanner	4	49
warranty	28	30	ohmic	15	49
aqueous	9	30	anodic	9	49
autonomously	9	30	comsol	2	48
limn	6	30	cycled	13	48
latch	3	30	indiana	2	48
illuminated	4	30	correction	16	48
redundant	8	30	automation	20	47
photocurrent	8	30	diode	11	47
ethanol	10	30	equalization	4	47
biped	2	30	defect	15	47
leaks	2	30	upgrade	12	47
audit	9	29	binder	14	47

Anlage A 7b

Beispiele von 25 Termini jeder Frequenzklasse aus dem Korpus WissPubl

Frequenzklasse: zwischen 51-80			Frequenzklasse: zwischen 81-100		
Terminus	**Dokument-frequenz**	**Wort-frequenz**	**Terminus**	**Dokument-frequenz**	**Wort-frequenz**
schematic	32	80	lattice	21	97
amplitude	9	80	solvent	17	97
width	26	80	rechargeable	12	96
lunar	2	80	turbine	11	95
charge-discharge	9	79	morphology	18	94
mesh	9	79	concentrator	8	93
tun	19	79	template	10	93
nanorod	4	79	inlet	10	93
sect	7	79	axis	27	93
june	27	78	carbonate	20	92
modell	11	78	illumination	13	92
operational	33	78	converter	7	92
separator	17	78	pre-cooling	4	91
dioxide	14	78	iii-v	10	90
pivot	3	75	collision	11	88
propulsion	7	75	conditioning	31	88
workpiece	2	74	ionic	16	88
dependence	19	74	entropy	4	88
differential	23	73	alloy	15	88
materials	19	73	diameter	25	88
combustion	22	73	wafer	9	87
utilization	27	73	greenhouse	17	87
generator	18	72	thereof	28	87
ambient	21	72	perl	2	87
incremental	15	72	membrane	13	87
galvanostatic	12	71	fabricate	26	86

Anlage A 7c

Beispiele von 25 Termini jeder Frequenzklasse aus dem Korpus WissPubl

Frequenzklasse: zwischen 101-150			Frequenzklasse: zwischen 151-200		
Terminus	**Dokument-frequenz**	**Wort-frequenz**	**Terminus**	**Dokument-frequenz**	**Wort-frequenz**
probability	8	150	appendix	18	199
bandgap	14	150	simulated	35	197
roadmap	4	149	abstraction	8	191
graphite	19	148	occupancy	18	191
torque	19	146	polymer	19	190
deployment	16	146	systems	56	188
gainp	12	145	composite	21	188
lithiation	4	144	catalyst	10	181
pollutant	15	139	lithium-ion	21	179
cooling	30	138	detection	21	179
robust	30	138	indoor	33	176
lbnl	13	135	attic	9	176
methodology	47	134	conductivity	30	174
nanoparticle	10	134	leak	10	174
mcmb	4	134	on-site	12	172
impedance	16	133	semiconductor	23	169
adaptive	25	128	thesis	13	169
manipulator	16	126	baseline	25	164
monolithic	9	125	simulator	15	164
predicate	2	125	porosity	9	163
emitter	7	124	criterion	39	157
summarize	49	121	concentrating	4	157
additive	18	120	tandem	13	157
coating	22	119	waypoint	2	156
wherein	13	119	http	32	154
printer	2	118	calibrate	19	154

Anlage A 7d

Beispiele von Termini jeder Frequenzklasse aus dem Korpus WissPubl

Frequenzklasse: zwischen 201-300			Frequenzklasse: zwischen 301-400		
Terminus	**Dokument-frequenz**	**Wort-frequenz**	**Terminus**	**Dokument-frequenz**	**Wort-frequenz**
payload	19	298	heating	70	388
modeling	57	295	plenum	9	386
anode	24	286	junction	22	343
insulation	21	279	module	30	337
coefficient	43	278	substrate	22	317
furnace	14	268	cathode	22	314
diffusion	28	268	handler	12	305
cells	25	260	optimal	47	304
velocity	21	249	trajectory	14	302
corresponding	59	246			
absorption	21	234			
graphene	7	234	**Frequenzklasse: zwischen 401-500**		
subcell	10	232	electrochemical	30	487
controll	64	228	leakage	17	460
cycling	28	226	savings	56	457
gaas	19	222	controller	40	452
hvac	32	218	robotic	28	409
calibration	10	208			
electron	28	206	**Frequenzklasse: zwischen 501-1000**		
porous	18	204	lithium	35	854
optimization	48	204	simulation	64	819
specification	36	201	duct	18	721
			parameter	74	696
			electrolyte	29	598
			silicon	16	573
			ventilation	30	533

Anlage A 8

Gruppierung der 181 in den beiden Analysekorpora enthaltenen Fachtermini nach Frequenzklassen

Frequenzgruppen	Anzahl Wörter	Je ein Beispiel für die niedrigste bzw. höchste Frequenz des Wortes in der jeweiligen Gruppe
bis 5	20	upside (beide Korpora: 2), millimeter (WissPubl: 5, TranskriptVorl: 15)
zwischen 6 und 9	24	unclear (WissPubl: 6, TranskriptVorl: 3), parabolic (WissPubl: 9, TranskriptVorl: 4)
zwischen 10 und 19	29	perpendicular (WissPubl: 10, TranskripVorl: 5); analyse (WissPubl: 19, TranskriptVorl: 2)
zwischen 20 und 29	15	encompass (WissPubl: 22, TranskriptVorl: 2); multiply (WissPubl: 29, TranskriptVorl: 27)
zwischen 30 und 39	18	emit (WissPubl: 30, TranskriptVorl: 2); displacement (WissPubl: 39, TranskriptVorl: 10)
zwischen 40 und 60	12	integral (WissPubl: 40, TranskriptVorl: 13); negligible (WissPubl: 53, TranskriptVorl: 3)
zwischen 61 und 80	12	conductor (WissPubl: 61, TranskriptVorl: 16); optimum (WissPubl: 79, TranskriptVorl: 6)
zwischen 81 und 100	12	zinc (WissPubl: 81, TranskriptVorl: 20); optical (WissPubl: 97, TranskriptVorl: 9)
zwischen 101 und 130	10	parallel (WissPubl: 101, TranskriptVorl: 15); smaller (WissPubl: 125, TranskriptVorl: 14)
zwischen 131 und 200	12	Photon (WissPubl: 133, TranskriptVorl: 11); Prototype (WissPubl: 188, TranskriptVorl: 2)
zwischen 201 und 300	6	oxide (WissPubl: 217, TranskriptVorl: 7); zero (WissPubl: 297, TranskriptVorl: 78);
zwischen 301 und 500	3	thickness (WissPubl: 318, TranskriptVorl: 26), algorithm (WissPubl: 413, TranskriptVorl: 2)
zwischen 501 und 1000	7	interface (WissPubl: 522, TranskriptVorl: 18), thermal (WissPubl: 701, TranskriptVorl: 52)
und darüber	1	datum (WissPubl: 1401; TranskriptVorl: 25)

Anlage A 9

181 Termini, die in beiden Analysekorpora enthalten sind

		WissPubl		TranskriptVorl	
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz	Dokument-frequenz
1	datum	127	1401	25	3
2	thermal	93	701	52	3
3	renewable	58	656	18	2
4	voltage	51	603	88	8
5	discharge	32	593	37	5
6	higher	99	575	29	8
7	configuration	53	560	5	2
8	interface	53	522	18	3
9	algorithm	46	413	2	2
10	hybrid	35	368	4	2
11	thickness	36	318	26	2
12	zero	58	297	78	8
13	grid	42	297	7	2
14	grind	7	274	14	2
15	accord	57	231	5	4
16	states	48	229	2	2
17	oxide	29	217	7	2
18	prototype	28	188	2	2
19	photovoltaic	38	175	8	2
20	node	15	169	4	2
21	inch	17	165	3	2
22	hydrogen	18	160	17	4
23	correspond	49	158	4	2
24	static	27	148	9	3
25	linear	51	144	10	2
26	institute	53	142	3	2
27	matrix	30	140	2	2
28	ion	23	135	28	2
29	photon	15	133	11	2
30	smaller	50	125	14	6
31	nominal	22	124	3	2
32	copper	14	124	24	3
33	percent	25	122	28	3

		WissPubl		TranskriptVorl	
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz	Dokument-frequenz
34	computation	11	121	2	2
35	diagram	24	120	5	3
36	terminal	14	118	3	3
37	geometry	20	116	8	3
38	compute	14	105	15	2
39	parallel	33	101	15	4
40	optical	27	97	9	2
41	means	50	96	6	2
42	capacitance	14	94	41	3
43	geothermal	9	94	4	2
44	verify	40	91	10	2
45	respective	16	90	3	2
46	radius	19	87	27	4
47	optimized	35	85	2	2
48	vector	21	85	3	3
49	optimize	48	82	6	2
50	insulate	14	81	4	2
51	zinc	3	81	20	2
52	optimum	20	79	6	3
53	equilibrium	18	77	20	2
54	capacitor	17	74	116	3
55	excess	26	70	2	2
56	plug	12	69	6	4
57	extract	28	68	2	2
58	simplify	32	68	2	2
59	seconds	18	68	12	5
60	fossil	16	63	8	2
61	overview	26	63	4	2
62	foil	15	63	2	2
63	conductor	11	61	16	2
64	negligible	25	53	3	2
65	subsystem	17	53	6	2
66	efficiently	29	47	2	2

Anlage A 9a

181 Termini, die in beiden Analysekorpora enthalten sind

		WissPubl		TranskriptVorl	
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz	Dokument-frequenz
67	inverter	4	47	40	2
68	finite	12	47	3	2
69	compensate	25	45	9	2
70	watt	12	43	58	4
71	graph	15	43	11	2
72	precision	14	42	5	2
73	approximate	12	42	3	2
74	horizontal	19	42	3	2
75	integral	17	40	13	4
76	displacement	13	39	10	2
77	mile	11	39	8	2
78	supplies	16	38	2	2
79	mobility	16	38	2	2
80	decay	14	37	4	2
81	namely	20	37	2	2
82	exponential	12	37	6	3
83	rotate	11	35	3	2
84	numerical	13	35	2	2
85	atomic	14	35	2	2
86	ton	12	33	2	2
87	proportional	18	32	3	2
88	papers	15	32	2	2
89	independently	20	32	4	3
90	infrared	10	32	5	3
91	warming	10	31	2	2
92	instantaneous	14	31	7	2
93	emit	12	30	2	2
94	multiply	20	29	27	7
95	bonding	5	29	3	2
96	simplicity	19	28	2	2
97	lightweight	9	27	9	2
98	substitute	17	27	16	7
99	atom	12	27	3	2

		WissPubl		TranskriptVorl	
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz	Dokument-frequenz
100	complicate	19	27	9	3
101	converge	7	25	2	2
102	cent	7	23	4	3
103	consuming	16	23	9	3
104	standby	4	22	21	2
105	fuse	5	22	14	2
106	transparency	5	22	2	2
107	encompass	15	22	2	2
108	straightforward	19	22	5	2
109	analyse	8	19	2	2
110	individually	11	19	2	2
111	dilute	7	19	7	2
112	molten	3	18	6	2
113	coulomb	5	17	4	2
114	resistor	8	17	44	3
115	exponentially	6	17	2	2
116	faraday	6	16	7	2
117	cubic	9	16	3	3
118	reactor	5	16	3	2
119	iterative	10	15	6	3
120	laptop	10	15	6	3
121	electrostatic	4	15	6	2
122	older	10	15	3	2
123	overhead	7	14	3	2
124	slower	12	14	5	4
125	illustrative	9	14	2	2
126	energetic	6	13	2	2
127	whilst	5	13	2	2
128	volt	7	13	79	5
129	minus	7	13	59	8
130	fastest	8	12	3	2
131	cheaper	9	12	6	2
132	focal	5	11	8	2

Anlage A 9b

181 Termini, die in beiden Analysekorpora enthalten sind

		WissPubl		TranskriptVorl	
Rang	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz	Dokument-frequenz
133	polarity	3	10	5	2
134	grazing	2	10	3	2
135	perpendicular	7	10	2	2
136	tradeoff	9	10	4	2
137	mature	7	10	5	2
138	parabolic	4	9	4	2
139	dissipation	3	9	9	2
140	analogy	3	9	5	3
141	kelvin	4	9	18	3
142	sketch	4	8	5	2
143	convection	5	8	15	2
144	right-hand	7	8	2	2
145	demo	4	8	2	2
146	deflection	3	8	2	2
147	dissipate	4	7	37	4
148	triangle	4	7	2	2
149	welding	3	7	2	2
150	physicist	2	7	4	2
151	nowadays	4	7	6	4
152	suffice	6	7	3	2
153	orbital	4	7	11	2
154	kilowatt	5	7	9	3
155	interestingly	6	7	3	3
156	lump	4	6	3	2
157	fundamentally	6	6	3	2
158	whereby	3	6	6	3
159	weld	4	6	5	2
160	cylinder	4	6	5	3
161	unclear	4	6	3	2
162	millimeter	3	5	15	3
163	elliptical	3	5	4	3
164	joule	2	5	50	3
165	ray	3	5	4	2
166	kilogram	3	5	9	2
167	spontaneously	4	5	2	2
168	kilowatt-hour	4	4	2	2
169	iterate	2	4	3	2
170	suitably	4	4	2	2
171	fahrenheit	3	4	2	2
172	infinity	4	4	15	3
173	infinitely	3	4	4	3
174	clockwise	4	4	5	2
175	fumes	2	3	2	2
176	one-half	2	3	21	2
177	sulfuric	2	3	5	2
178	microphone	3	3	2	2
179	uranium	2	3	23	3
180	maturity	2	2	2	2
181	upside	2	2	2	2

Anlage A 10

Die 26 gemeinsamen Termini in den Analysekorpora, die der zweiten Wortformklasse angehören und einen Häufigkeitsquotienten über 10 haben

	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz in COCA	Relative Häufigkeit in WissPubl	Relative Häufigkeit in COCA	Häufigkeits-quotient
1	energy	196	5573	64139	6,06E-03	1,43E-04	42,489
2	potential	110	832	31289	9,04E-04	6,95E-05	13,003
3	battery	50	1405	9793	1,53E-03	2,18E-05	70,157
4	circuit	36	278	6747	3,02E-04	1,50E-05	20,149
5	solar	57	2011	12121	2,19E-03	2,69E-05	81,131
6	temperature	86	1239	23111	1,35E-03	5,14E-05	26,216
7	resistance	56	528	15686	5,74E-04	3,49E-05	16,460
8	surface	63	977	39367	1,06E-03	8,75E-05	12,136
9	electric	73	585	12007	6,36E-04	2,67E-05	23,825
10	supply	68	796	24920	8,65E-04	5,54E-05	15,620
11	constant	68	495	14069	5,38E-04	3,13E-05	17,205
12	load	88	883	10736	9,60E-04	2,39E-05	40,219
13	calculate	74	331	10485	3,60E-04	2,33E-05	15,437
14	equation	52	445	7372	4,84E-04	1,64E-05	29,518
15	flow	56	844	18469	9,17E-04	4,10E-05	22,347
16	transfer	59	332	14686	3,61E-04	3,26E-05	11,055
17	efficiency	169	1571	10579	1,71E-03	2,35E-05	72,618
18	receiver	7	507	8882	5,51E-04	1,97E-05	27,913
19	concentration	37	643	12146	6,99E-04	2,70E-05	25,888
20	electricity	68	632	11954	6,87E-04	2,66E-05	25,853
21	output	58	306	7041	3,33E-04	1,56E-05	21,252
22	calculation	51	187	5577	2,03E-04	1,24E-05	16,397
23	input	53	297	7004	3,23E-04	1,56E-05	20,736
24	integration	46	290	8847	3,15E-04	1,97E-05	16,029
25	loop	27	158	6175	1,72E-04	1,37E-05	12,512
26	equivalent	49	210	5935	2,28E-04	1,32E-05	17,303

Anlage A 11

Ermittlung von potentiellen fachspezifischen Termini in der 2. Wortformenklasse von WissPubl mittels Häufigkeitsquotienten

Terminus	DF	TF	SVerweis	TF in COCA	RelFreq in WissPubl	RelFrreq in COCA	HQ
robot	36	3495	6023	6023	3,80E-03	1,34E-05	283,757
sensor	42	1111	5154	5154	1,21E-03	1,15E-05	105,410
solar	57	2011	12121	12121	2,19E-03	2,69E-05	81,131
efficiency	169	1571	10579	10579	1,71E-03	2,35E-05	72,618
battery	50	1405	9793	9793	1,53E-03	2,18E-05	70,157
density	55	618	6040	6040	6,72E-04	1,34E-05	50,034
exploration	14	663	7294	7294	7,20E-04	1,62E-05	44,449
energy	196	5573	64139	64139	6,06E-03	1,43E-04	42,489
cell	74	3722	44831	44831	4,04E-03	9,96E-05	40,599
load	88	883	10736	10736	9,60E-04	2,39E-05	40,219
sustainable	32	428	5339	5339	4,65E-04	1,19E-05	39,201
measurement	79	669	10187	10187	7,27E-04	2,26E-05	32,114
update	25	339	5227	5227	3,68E-04	1,16E-05	31,715
carbon	60	675	10799	10799	7,34E-04	2,40E-05	30,566
maximum	66	493	8075	8075	5,36E-04	1,79E-05	29,855
equation	52	445	7372	7372	4,84E-04	1,64E-05	29,518
consumption	88	600	10162	10162	6,52E-04	2,26E-05	28,873
receiver	7	507	8882	8882	5,51E-04	1,97E-05	27,913
temperature	86	1239	23111	23111	1,35E-03	5,14E-05	26,216
concentration	37	643	12146	12146	6,99E-04	2,70E-05	25,888
electricity	68	632	11954	11954	6,87E-04	2,66E-05	25,853
particle	32	404	7925	7925	4,39E-04	1,76E-05	24,928
electrical	82	367	7347	7347	3,99E-04	1,63E-05	24,427
curve	49	415	8397	8397	4,51E-04	1,87E-05	24,168
layer	54	794	16250	16250	8,63E-04	3,61E-05	23,894
electric	73	585	12007	12007	6,36E-04	2,67E-05	23,825
experimental	67	427	9019	9019	4,64E-04	2,00E-05	23,152
flow	56	844	18469	18469	9,17E-04	4,10E-05	22,347
cycle	59	698	15775	15775	7,59E-04	3,51E-05	21,637
output	58	306	7041	7041	3,33E-04	1,56E-05	21,252
input	53	297	7004	7004	3,23E-04	1,56E-05	20,736
capacity	74	887	21102	21102	9,64E-04	4,69E-05	20,555
storage	57	457	10893	10893	4,97E-04	2,42E-05	20,515
circuit	36	278	6747	6747	3,02E-04	1,50E-05	20,149
obstacle	22	329	8075	8075	3,58E-04	1,79E-05	19,924
constraint	35	239	5942	5942	2,60E-04	1,32E-05	19,669
laser	12	287	7197	7197	3,12E-04	1,60E-05	19,500
respectively	72	317	8223	8223	3,44E-04	1,83E-05	18,851
peak	58	558	14481	14481	6,06E-04	3,22E-05	18,843
motor	26	384	9977	9977	4,17E-04	2,22E-05	18,821
operator	29	350	9351	9351	3,80E-04	2,08E-05	18,303
emission	51	420	11688	11688	4,56E-04	2,60E-05	17,572
equivalent	49	210	5935	5935	2,28E-04	1,32E-05	17,303
constant	68	495	14069	14069	5,38E-04	3,13E-05	17,205
transport	38	187	5336	5336	2,03E-04	1,19E-05	17,137
application	119	899	25735	25735	9,77E-04	5,72E-05	17,082
resistance	56	528	15686	15686	5,74E-04	3,49E-05	16,460
calculation	51	187	5577	5577	2,03E-04	1,24E-05	16,397
integration	46	290	8847	8847	3,15E-04	1,97E-05	16,029
coordinate	27	167	5152	5152	1,81E-04	1,14E-05	15,851
register	14	304	9389	9389	3,30E-04	2,09E-05	15,833
planning	35	528	16434	16434	5,74E-04	3,65E-05	15,711
supply	68	796	24920	24920	8,65E-04	5,54E-05	15,620
chapter	14	334	10579	10579	3,63E-04	2,35E-05	15,439
calculate	74	331	10485	10485	3,60E-04	2,33E-05	15,437
architecture	33	266	8449	8449	2,89E-04	1,88E-05	15,395
collector	25	217	6976	6976	2,36E-04	1,55E-05	15,211
mode	45	369	12107	12107	4,01E-04	2,69E-05	14,904
decrease	71	282	9291	9291	3,06E-04	2,06E-05	14,842
design	131	1457	48605	48605	1,58E-03	1,08E-04	14,659
feedback	37	228	7735	7735	2,48E-04	1,72E-05	14,414
figure	85	2205	74858	74858	2,40E-03	1,66E-04	14,404
integrate	70	263	9060	9060	2,86E-04	2,01E-05	14,195

efficient	84	303	10452	10452	3,29E-04	2,32E-05	14,176
reduction	87	456	15862	15862	4,96E-04	3,52E-05	14,058
laboratory	72	285	9972	9972	3,10E-04	2,22E-05	13,976
ratio	65	298	10457	10457	3,24E-04	2,32E-05	13,935
filter	32	176	6210	6210	1,91E-04	1,38E-05	13,859
envelope	21	205	7241	7241	2,23E-04	1,61E-05	13,844
specify	41	152	5379	5379	1,65E-04	1,20E-05	13,818
capability	62	308	10906	10906	3,35E-04	2,42E-05	13,810
angle	24	295	10456	10456	3,21E-04	2,32E-05	13,797
device	65	701	24918	24918	7,62E-04	5,54E-05	13,757
minimum	55	225	8011	8011	2,45E-04	1,78E-05	13,734
mechanical	57	220	7835	7835	2,39E-04	1,74E-05	13,731
disk	10	221	7898	7898	2,40E-04	1,76E-05	13,683
scan	22	192	6879	6879	2,09E-04	1,53E-05	13,649
conversion	38	157	5627	5627	1,71E-04	1,25E-05	13,644
residential	37	205	7415	7415	2,23E-04	1,65E-05	13,519
summary	54	146	5304	5304	1,59E-04	1,18E-05	13,461
error	45	438	15980	15980	4,76E-04	3,55E-05	13,403
accuracy	55	188	6942	6942	2,04E-04	1,54E-05	13,243
model	120	2244	82973	82973	2,44E-03	1,84E-04	13,225
potential	110	832	31289	31289	9,04E-04	6,95E-05	13,003
flexible	35	167	6312	6312	1,81E-04	1,40E-05	12,938
mass	59	479	18107	18107	5,21E-04	4,02E-05	12,936
lower	86	453	17206	17206	4,92E-04	3,82E-05	12,875
suitable	53	139	5291	5291	1,51E-04	1,18E-05	12,847
fuel	42	622	23934	23934	6,76E-04	5,32E-05	12,708
phase	52	416	16045	16045	4,52E-04	3,57E-05	12,678
current	106	1583	61252	61252	1,72E-03	1,36E-04	12,638
lighting	25	131	5081	5081	1,42E-04	1,13E-05	12,608
invention	8	134	5211	5211	1,46E-04	1,16E-05	12,575
building	152	2011	78487	78487	2,19E-03	1,74E-04	12,529
installation	41	152	5938	5938	1,65E-04	1,32E-05	12,517
loop	27	158	6175	6175	1,72E-04	1,37E-05	12,512
attribute	32	236	9228	9228	2,56E-04	2,05E-05	12,506
experiment	55	470	18440	18440	5,11E-04	4,10E-05	12,464
plot	42	266	10484	10484	2,89E-04	2,33E-05	12,407
dynamic	44	179	7175	7175	1,95E-04	1,59E-05	12,200
surface	63	977	39367	39367	1,06E-03	8,75E-05	12,136
carrier	16	229	9251	9251	2,49E-04	2,06E-05	12,105
method	95	1048	42667	42667	1,14E-03	9,48E-05	12,011
stack	21	143	5858	5858	1,55E-04	1,30E-05	11,937
demand	90	741	30519	30519	8,05E-04	6,78E-05	11,873
profile	49	292	12089	12089	3,17E-04	2,69E-05	11,812
spectrum	29	173	7216	7216	1,88E-04	1,60E-05	11,724
obtain	86	585	24538	24538	6,36E-04	5,45E-05	11,658
stability	38	252	10901	10901	2,74E-04	2,42E-05	11,304
joint	31	356	15444	15444	3,87E-04	3,43E-05	11,272
measure	120	1006	43938	43938	1,09E-03	9,76E-05	11,196
evaluate	85	361	15780	15780	3,92E-04	3,51E-05	11,187
hardware	23	158	6951	6951	1,72E-04	1,54E-05	11,115
vehicle	45	668	29421	29421	7,26E-04	6,54E-05	11,103
comprise	32	142	6258	6258	1,54E-04	1,39E-05	11,096
transfer	59	332	14686	14686	3,61E-04	3,26E-05	11,055
estimate	81	462	20474	20474	5,02E-04	4,55E-05	11,034
fraction	28	119	5313	5313	1,29E-04	1,18E-05	10,953
initial	77	424	19019	19019	4,61E-04	4,23E-05	10,902
implementation	51	207	9293	9293	2,25E-04	2,07E-05	10,892
channel	17	285	12975	12975	3,10E-04	2,88E-05	10,741
utility	50	286	13150	13150	3,11E-04	2,92E-05	10,635
illustrate	68	328	15091	15091	3,56E-04	3,35E-05	10,628
mobile	23	116	5339	5339	1,26E-04	1,19E-05	10,625
tolerance	19	119	5513	5513	1,29E-04	1,23E-05	10,555
dynamics	22	151	7084	7084	1,64E-04	1,57E-05	10,423
performance	158	1273	59909	59909	1,38E-03	1,33E-04	10,391
exhaust	13	110	5204	5204	1,20E-04	1,16E-05	10,336
task	42	812	39209	39209	8,82E-04	8,71E-05	10,127
autonomy	12	120	5849	5849	1,30E-04	1,30E-05	10,033

Anlage 12

Ermittlung von potentiellen fachspezifischen Termini in der 2. Wortformenklasse von TranskriptVorl mittels Häufigkeitsquotienten

	Terminus	Dokument-frequenz	Wort-frequenz	Wort-frequenz in COCA	Relative Häufigkeit in TranskriptVorl	Relative Häufigkeit in COCA	Häufigkeits-quotient
1	circuit	9	85	6747	1,10E-03	1,50E-05	73,07
2	battery	5	94	9793	1,21E-03	2,18E-05	55,67
3	bulb	3	47	5157	6,06E-04	1,15E-05	52,86
4	orbit	2	56	6490	7,22E-04	1,44E-05	50,04
5	square	8	88	11630	1,13E-03	2,58E-05	43,88
6	solar	3	76	12121	9,80E-04	2,69E-05	36,36
7	electric	3	66	12007	8,51E-04	2,67E-05	31,88
8	potential	6	167	31289	2,15E-03	6,95E-05	30,96
9	lecture	9	44	8278	5,67E-04	1,84E-05	30,83
10	equation	7	39	7372	5,03E-04	1,64E-05	30,68
11	resistance	5	72	15686	9,28E-04	3,49E-05	26,62
12	energy	9	279	64139	3,60E-03	1,43E-04	25,23
13	switch	6	54	12508	6,96E-04	2,78E-05	25,04
14	calculate	5	44	10485	5,67E-04	2,33E-05	24,34
15	load	4	44	10736	5,67E-04	2,39E-05	23,77
16	charge	7	168	42353	2,17E-03	9,41E-05	23,01
17	structure	4	173	43638	2,23E-03	9,70E-05	22,99
18	constant	7	51	14069	6,57E-04	3,13E-05	21,02
19	calculation	3	20	5577	2,58E-04	1,24E-05	20,80
20	plate	5	88	24592	1,13E-03	5,46E-05	20,75
21	meter	6	35	10246	4,51E-04	2,28E-05	19,81
22	divide	8	62	18337	7,99E-04	4,07E-05	19,61
23	shuttle	2	25	7593	3,22E-04	1,69E-05	19,10
24	radiation	5	26	7912	3,35E-04	1,76E-05	19,06
25	earth	4	70	21350	9,02E-04	4,74E-05	19,02
26	tension	2	51	15614	6,57E-04	3,47E-05	18,94
27	equal	9	64	19952	8,25E-04	4,43E-05	18,60
28	spark	2	16	5071	2,06E-04	1,13E-05	18,30
29	temperature	7	72	23111	9,28E-04	5,14E-05	18,07
30	mirror	2	62	20081	7,99E-04	4,46E-05	17,91
31	cable	2	56	18277	7,22E-04	4,06E-05	17,77
32	liquid	4	18	5948	2,32E-04	1,32E-05	17,55
33	dish	2	57	18887	7,35E-04	4,20E-05	17,50
34	output	3	21	7041	2,71E-04	1,56E-05	17,30
35	receiver	2	26	8882	3,35E-04	1,97E-05	16,98
36	consume	3	30	10308	3,87E-04	2,29E-05	16,88
37	frequency	3	32	11284	4,12E-04	2,51E-05	16,45
38	gate	2	50	17999	6,44E-04	4,00E-05	16,11

39	input	4	19	7004	2,45E-04	1,56E-05	15,73
40	plus	8	46	17202	5,93E-04	3,82E-05	15,51
41	efficiency	3	26	10579	3,35E-04	2,35E-05	14,25
42	heat	7	94	40363	1,21E-03	8,97E-05	13,51
43	loop	4	14	6175	1,80E-04	1,37E-05	13,15
44	transfer	3	32	14686	4,12E-04	3,26E-05	12,64
45	concentration	4	26	12146	3,35E-04	2,70E-05	12,42
46	flow	6	38	18469	4,90E-04	4,10E-05	11,93
47	supply	5	51	24920	6,57E-04	5,54E-05	11,87
48	equivalent	4	12	5935	1,55E-04	1,32E-05	11,73
49	clock	3	25	12395	3,22E-04	2,75E-05	11,70
50	power	9	277	141357	3,57E-03	3,14E-04	11,36
51	electricity	4	23	11954	2,96E-04	2,66E-05	11,16
52	sphere	2	11	5881	1,42E-04	1,31E-05	10,85
53	engineer	3	33	17991	4,25E-04	4,00E-05	10,64
54	reaction	5	49	26771	6,32E-04	5,95E-05	10,62
55	surface	7	72	39367	9,28E-04	8,75E-05	10,61
56	integration	3	16	8847	2,06E-04	1,97E-05	10,49
57	signal	4	30	16835	3,87E-04	3,74E-05	10,34

Anlagen B

Dokumentation der Rechercheinstrumente

Anlage B1

Übersicht über die verwendeten Rechercheinstrumente

DB-Bezeichnung	Datenbanktyp							Zugänglichkeit	
	Fach-bibliografie	Interdisziplin. Bibliografie	Aufsatz-bibliografie	Fach-Portal	Verlags-portal	Abstracts	Volltexte	Lizenz-pflichtig	Lizenzfrei
ENTEC http://www.wti-frankfurt.de/images/stories/download/de-etec.pdf	☑					☑	★	☑	
ETDEWEB https://www.etde.org/etdeweb/	☑						☑		☑
GetInfo https://getinfo.de/app/				☑					☑ ★
INSPEC http://www.theiet.org/resources/inspec/	☑					☑		☑	
SciVerse Hub (Scopus, ScienceDirect) http://www.hub.sciverse.com/action/home		☑	☑		☑	☑	☑		☑ ★
SpringerLink http://link.springer.com/		☑	☑		☑	☑	☑		☑ ★
TEMA http://www.wti-frankfurt.de/images/stories/download/de-tema.pdf	☑					☑			☑
ViFaTec http://www.vifatec.de/				☑		☑	☑		☑ ★
Web of Science http://thomsonreuters.com/products_services/science/science_products/a-z/web_of_science/		☑	☑			☑	☑	☑	
Willey http://onlinelibrary.wiley.com/			☑				☑		☑ ★
Erklärung zu ★									
★=Die Recherche ist kostenlos, der Zugang zu einzelnen Publikationen kann kostenpflichtig sein									
★★=Volltexte werden in Ausnahmefällen angeboten									

Anlage B1a

Übersicht über die verwendeten Rechercheinstrumente - Suchmaschinen

Bezeichnung	**URL**
Google Scholar	http://scholar.google.de/
Scirus	http://www.scirus.com/
BASE Bielefeld Academic Search Engine	http://base.ub.uni-bielefeld.de
DOAJ Directory of Open Access Journals	http://www.doaj.org/
Ixquick Search Engine	https://www.ixquick.com/
Pandia Metasearch Engine	http://www.pandia.com/metasearch/index.html

Anlagen C

Domänen-Training und Adaption des Sprachmodells der automatischen Spracherkennung

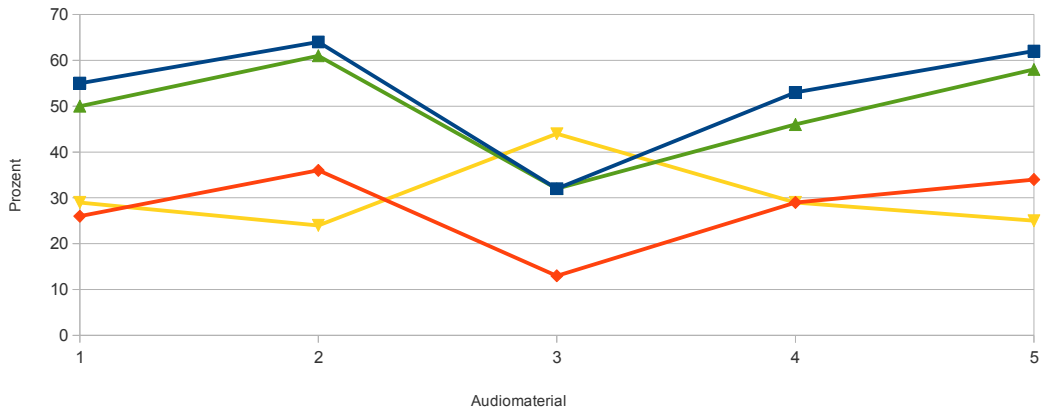
TIB | Technische Informationsbibliothek

Autor: Ingo Rieger

Version: 1.2
Angelegt am: 27.03.2012
Geändert am: 30.03.2012

Untersuchtes Audiomaterial und Bewertung der automatischen Spracherkennung des Produkts *Vetail-X* (European Media Laboratory) im Kontext des AV-Portals

Nr.	Sprache	Youtube-Titel	Youtube-Link	Originaltext	Erkannter Text (automatische Transkription mit Standard-sprachmodell Englisch, ohne Domaintraining)	Erkennungsgenauigkeit in Prozent nach EML-Angaben (100% – Prozentsatz falsch erkannter Wörter)	Erkennungsgenauigkeit in Prozent nach HPI-Tool (100% – Prozentsatz falsch erkannter Wörter)	BLEU Score (*) in Prozent nach TIB ITE Analysen	Levenshtein Distance (**) in Prozent nach TIB ITE Analysen
1	EN	Online Chemistry Lecture - Relative Melting Point	http://www.youtube.com/watch?v=Cqy6l8jZCdg	Oh, hi. Up on this board I've gone ahead and compartmentalized some compounds. What we have on the far left are the inert gases,group	hi up on the photo I've got a hidden compartments some compounds what we have on the father the unit gases the group greeting such as	55	50	26	29
2	EN	Using enzymes to generate hydrogen	http://www.youtube.com/watch?v=dR0lg43BNvU	I'm Roger Ely. And here at OSU I'm an Associate Professor in the department of Biological and Ecological Engineering. Starting about seven or eight	yeah I'm marginally and um room here it always you I'm an associate professor in the department biological anthropological engineering	64	61	36	24
3	EN	Converting Waste Water Into Energy	http://www.youtube.com/watch?v=h4sr0B2b3pg	My name is Hong Liu. I'm assistant professor at Oregon State University. I'm currently working on microbial fused cell based technologies:	many scenario I must confess Orenstein risky unfortunately my forgive yourself please technologies tragedy energy foundlings what the mark of if he	32	32	13	44
4	EN	Sustainable Energy: Wind Energy	http://www.youtube.com/watch?v=fySF5yVPNiY	My name is Ted Brekken. I'm a professor of electrical engineering at Oregon State University. There's a limited amount of this variable	i time instead 10 reckon number of astronautical engineering or against a university is said there is the limited amounts of this variable generation such	53	46	29	29
5	EN	Diode Tutorial & How to build an AC to DC power supply	http://www.youtube.com/watch?v=cyhzpFqXwdA	In this tutorial, I'm going to cover silicon diodes, bridge rectifiers, and how you convert AC to DC. Here's the schematic symbol of a diode and a	in this to try and going to cover silicon doubts rejected fires in how you convert 82 DC here's the schematic symbol of the dialogue and a picture of the	62	58	34	25



Anmerkungen:

(*) BLEU Score gemessen mit Implementierung *lingutil r2* unter http://code.google.com/p/lingutil/

(**) Levenshtein Distance gemessen mit Implementierung *Apache Commons Lang 3.1* unter http://commons.apache.org/lang/

Zur Interpretation der Werte:

Ein BLEU Score (Bilingual Evaluation Understudy) von 33% würde aussagen, dass ein Mensch die Ähnlichkeit zwischen transkribiertem Text und Originaltext in seiner Gesamtheit im Schnitt als 1/3 bewerten würde (s.a. http://en.wikipedia.org/wiki/BLEU).

Eine Levenshtein Distance von 33% würde aussagen, dass rund ein Drittel der Zeichen des Originaltextes in der transkribierten Version vertauscht, eingefügt oder entfernt worden sind (s.a. http://en.wikipedia.org/wiki/Levenshtein_distance).

Anlage C2

Trainings- und Testdaten für die Adaption des Sprachmodells der automatischen Spracherkennung

Projekt-bezeichnung	**Datum**	**Trainingsdaten**			
		Webdaten in Webseiten	Text-dokumente	Trans-kriptionen	Anzahl Wörter
Computer Science	16.11.12	11	4		447.407
	22.11.12	26	2	2	447.858
	10.12.12	27	6	4	461.273
		64	***12***	***6***	***1.356.538***
Technology	23.11.12	5	1		463.182
	25.11.12	19			465.774
	14.12.12	22	5	12	448.001
		46	***6***	***12***	***1.376.957***

Projekt-bezeichnung	**Datum**	**Testdaten**			
		Audiodatei + Transkription	Nur Transkript	Anzahl Minuten	Anzahl Wörter
Computer Science	04.12.12	1	1	429	
	08.12.12	2			
	11.12.12	7			
		10	***1***	***429***	***130.378***
Technology	16.12.12	7		110	18.873
	20.12.12	5		304,52	41.920
		12	***1***	***414,52***	***60.793***

Anlage C3

Trainingsdaten für das Fach Informatik in Form von Transkriptionen

Institution	Typ	Reihe	Nummer Reihe / Vorlesung	Titel	Sprecher	Semester	Datum	Dauer	URL	Lizenz
Harvard Extension School	Lecture	Understanding Computers and the Internet	E-1 / Lecture 4	The Internet	Malan, David	Herbst 2006	11.10.06	01:44:52	http://cdn.computerscience1.net/2006/fall/lectures/4/transcript4.pdf?download	Creative Commons - CC BY-NC-SA 3.0
Harvard Extension School	Lecture	Understanding Computers and the Internet	E-1 / Lecture 8	Security	Malan, David	Herbst 2006	k.A.	01:45:55	http://cdn.computerscience1.net/2006/fall/lectures/8/transcript8.pdf?download	Creative Commons - CC BY-NC-SA 3.0
Falling Walls International Conference on Future Breakthroughs in Science and Society	Speech	Falling Walls Conference 2011		Breaking the Wall of Data Deluge. How Efficient Data Exploration Enables New Scientific Discoveries	Ailamaki, Anastasia		2011	00:13:17	http://falling-walls.com/lectures/anastasia-ailamaki/	frei zugänglich als PDF/Html, keine weitere Angabe
Falling Walls International Conference on Future Breakthroughs in Science and Society	Speech	Falling Walls Conference 2011		Breaking the Wall of Computer Stupidity. How Wavelet Analysis Improves Geophysics, Biology and Art History	Daubechies, Ingrid		2011	00:16:36	http://falling-walls.com/lectures/ingrid-daubechies/transcript/	frei zugänglich als PDF/Html, keine weitere Angabe
Interarbor Solutions, LLC	Podcast			A Technical Look at How Parallel Processing Brings Vast New Capabilities to Large-Scale Data Analysis	Dana Gardner		05.01.09	k.A.	http://briefingsdirect.blogspot.de/2009/01/technical-look-at-how-parallel.html	Copyright Interarbor Solutions, LLC, 2005-2009. All rights reserved.
WAMU 88.5 FM American University Radio	Podcast			A Shift in Cloud Computing	k.A.		k.A.	12:58:17	http://thekojonnamdishow.org/shows/2011-06-14/shift-cloud-computing/transcript	Transcripts of WAMU programs are available for personal use.

Anlage C4

Testdaten für das Fach Informatik

Themen: Computer Science versus Programming & Software Engineering

	Sprecher	Titel	Typ	Reihe	Reihennr.	Datum	URL	Min.
1.	Edward Felten & Larry Peterson & Shirley Tilghman	Rip, Mix, Burn, Sue: Technology, Politics, and the Fight to Control Digital Media; Transkript von Carlos Ovalle	Lecture	Princeton University President's Lecture Series ⓘ	Nr. 1, 2004/05	12.10.04	http://www.cs.princeton.edu/~felten/rip/ Transkript: http://www.ischool.utexas.edu/~i312co/copyright/felten.html	76
2.	Sahami, Mehran	Programming Methodology-Lecture01	Lecture	computer science - Programming Methodology ⓘ	Course Meetings: 28	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture01.pdf	50
3.	Sahami, Mehran	Programming Methodology-Lecture02	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture02.pdf	48
4.	Sahami, Mehran	Programming Methodology-Lecture03	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture03.pdf	53
5.	Sahami, Mehran	Programming Methodology-Lecture04	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture04.pdf	51
6.	Sahami, Mehran	Programming Methodology-Lecture05	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture05.pdf	52
7.	Sahami, Mehran	Programming Methodology-Lecture06	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture06.pdf	50
8.	Sahami, Mehran	Programming Methodology-Lecture24	Lecture	s.o.	s. o.	k.A.	http://see.stanford.edu/materials/icspmcs106a/transcripts/ProgrammingMethodology-Lecture24.pdf	49
								429

Anlage C5

Trainingsdaten für das Fach Informatik in Form von Transkriptionen zum Thema Energietechnik

Workshop-Diskussionen der California Energy Commission
URL: http://www.energy.ca.gov/2011_energypolicy/documents/#11142011

	Bezeichnung	Datum
1.	Staff Workshop on the Role of Alternative Fuels in California's Transportation Energy Future	14.11.11
2.	Joint Committee Workshop on Natural Gas Market Assessment Reference Case and Scenario Results	27.09.11
3.	Committee Workshop on Draft Renewable Power in California: Status and Issues	14.09.11
4.	Transportation Committee Workshop on Transportation Energy Demand and Fuel Infrastructure Requirements	09.09.11
5.	Committee Workshop on 2012-2022 Preliminary Staff Electricity and Natural Gas Demand Forecast	30.08.11
6.	Committee Workshop on California Nuclear Power Plant Issues	26.07.11
7.	Staff Workshop on Achieving Energy Savings in California Buildings	20.07.11
8.	Committee Workshop on the California Clean Energy Future	06.07.11
9.	Committee Workshop on Distribution Infrastructure Challenges and Smart Grid Solutions to Advance 12,000 Megawatts of Distributed Generation	22.06.11
10.	Committee Workshop on Renewable, Localized Generation	09.05.11
11.	Committee Workshop on Energy Storage for Renewable Integration	28.04.11
12.	Joint Committee Workshop on Electricity Infrastructure Need Assessment	23.11.10

Anlage C6

Testdaten für das Fach Technik zum Thema Energie - Test 1

	Sprecher	Titel	Typ	Institution	Reihe	Tokens	Datum	Dauer in Min.	URL	Lizenz
1.	Sadoway, Donald	The missing link to renewable energy	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	1921	26.03.12	15,15	http://www.ted.com/talks/donald_sadoway_the_missing_link_to_renewable_energy.html	CC BY-NC_ND 3.0
2.	La Grou, John	John La Grou plugs smart power outlets	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	551	09.06.09	4,1	http://www.ted.com/talks/john_la_grou_plugs_smart_power_outlets_1.html	CC BY-NC_ND 3.1
3.	Gross, Bill	Bill Gross on new energy	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	4484	02.02.09	19,53	http://www.ted.com/talks/bill_gross_on_new_energy.html	CC BY-NC_ND 3.2
4.	Griffith, Saul	High-altitude wind energy from kites!	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	913	23.03.09	5,53	http://www.ted.com/talks/saul_griffith_on_kites_as_the_future_of_renewable_energy.html	CC BY-NC_ND 3.3
5.	MacCready, Paul	Paul MacCready flies on solar wings	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	2766	26.09.07	21,24	http://www.ted.com/talks/paul_maccready_flies_on_solar_wings.html	CC BY-NC_ND 3.4
6.	Trent, Jonathan	Energy from floating algae pods	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	2532	08.09.12	15,16	http://www.ted.com/talks/jonathan_trent_energy_from_floating_algae_pods.html	CC BY-NC_ND 3.5
7.	Gates, Bill	Bill Gates on energy: Innovating to zero!	Rede	TED ⓘ	Ideas worth spreading from the TED Conference	4524	18.02.10	28,51	http://www.ted.com/talks/bill_gates.html	CC BY-NC_ND 3.6
						17691		**109,22**		

Anlage C6a

Testdaten für das Fach Technik zu den Themen Energie und Robotik - Test 2

	Sprecher	Titel	Typ	Institution	Reihe / Reihennr. / Vorlesungsnr.	Tokens	Datum	Dauer in Min.	URL
1.	Lewin, Walter	Orbits and Escape Velocity	Video Lecture	MIT ⓘ	Physics I / 8.01 Classical Mechanics / Lecture 14	7112	13.10.1999 (veröffentlicht am 03.01.08)	50,06	http://ocw.mit.edu/courses/physics/8-01-physics-i-classical-mechanics-fall-1999/video-lectures/lecture-14/
2.	Cummins, Christopher	Cell potentials and free energy	Video Lecture	MIT ⓘ	Principles of chemical science II / 5.112 / Lecture 23	5234	Herbst 2005 (veröffentlicht am 08.05.07)	43,08	http://ocw.mit.edu/courses/chemistry/5-112-principles-of-chemical-science-fall-2005/video-lectures/lecture-23-cell-potentials-and-free-energy/
3.	Lewin, Walter	Capacitance and Field Energy	Video Lecture	MIT ⓘ	Physics II / 8.02 Electricity & Magnetism /Lecture 7	7435	Frühjahr 2002 (veröffentlicht am 18.05.07)	49,26	http://ocw.mit.edu/courses/physics/8-02-electricity-and-magnetism-spring-2002/video-lectures/lecture-7-capacitance-and-field-energy/
4.	Lewin, Walter	Batteries and EMF	Video Lecture	MIT ⓘ	Physics II / 8.02 Electricity & Magnetism /Lecture 10	6715	Frühjahr 2002 (veröffentlicht am 18.05.07)	50,05	http://ocw.mit.edu/courses/physics/8-02-electricity-and-magnetism-spring-2002/video-lectures/lecture-10-batteries-and-emf/
5.	Lavoie, Anthony	Systems Engineering for Space Shuttle Payloads	Video Lecture	MIT ⓘ	Aircraft Systems Engineering / 16.885 J /Lecture 21	15430	Herbst 2005 (veröffentlicht am 09.06.09)	112,07	http://ocw.mit.edu/courses/aeronautics-and-astronautics/16-885j-aircraft-systems-engineering-fall-2005/video-lectures/lecture-21/#vid_related
						41926		**304,52**	

ⓘ MIT = Massachusetts Institute of Technology

Anlage C7

Evaluation der Adaption des Sprachmodells - OOV-Rate, WER und Perplexität
Fach Informatik

	Test-Set_ID		Words-in-Test	Known-Words	OOV-Rate %	OOV-Count	Perplexity	WER
1.	ComputerScience_EN_Felten	Adaptiertes Modell	13371	13312	0.44	59	339.402	78.3
	ComputerScience_EN_Felten	Basismodell	12722	12665	0.45	57	1177.76	88.6
2.	KIT_ComputerScience_EN	Adaptiertes Modell	12773	11307	11,48	1466	279.967	
	KIT_ComputerScience_EN	Basismodell	12481	11015	11,75	1466	1315.67	
3.	Stanford1a	Adaptiertes Modell	12174	12095	0.65	79	205.99	54.1
	Stanford1a	Basismodell	12170	12097	0.6	73	805.925	73.5
4.	Stanford2a	Adaptiertes Modell	11086	10954	1,19	132	254.441	55.0
	Stanford2a	Basismodell	11086	10954	1,19	132	992.217	74.0
5.	Stanford3	Adaptiertes Modell	12602	12484	0.94	118	293.286	59.3
	Stanford3	Basismodell	12606	12491	0.91	115	1536.36	75.9
6.	Stanford4	Adaptiertes Modell	10960	10864	0.88	96	367.151	54.5
	Stanford4	Basismodell	10967	10873	0.86	94	2233.05	75.7
7.	Stanford5	Adaptiertes Modell	11602	11470	1,07	132	354.782	59.0
	Stanford5	Basismodell	11635	11511	1,07	124	1863.62	76.1
8.	Stanford6	Adaptiertes Modell	11828	11698	1,05	130	300.994	59.2
	Stanford6	Basismodell	11854	11730	1,05	124	1408.09	74.7
9.	Stanford24	Adaptiertes Modell	11590	11498	0.79	92	192.84	53.0
	Stanford24	Basismodell	11602	11517	0.73	85	561.208	73.5
			215109	**210535**		**4574**		

Anlage C8

Evaluation der Adaption des Sprachmodells - OOV-Rate, WER und Perplexität

Fach Technik

	Test-Set	Modell	Words-in-Test	Known-Words	OOV-Rate %	OOV-Count	Perplexity	WCR	WER
1.	BatteriesEnergy	BM	7165	7115	0,7	50	329,159	51	51,8
	BatteriesEnergy	AM	7070	6986	1,19	84	258,657	62,6	40,6
2.	Biofuel	AM	448	441	1,56	7	381,616		
	Biofuel	BM	448	441					
3.	CellFreeEnergy	BM	5703	5592	1,95	111	235,198	65,6	42,7
	CellFreeEnergy	AM	5648	5492	2,76	156	184,811	73,2	37,7
4.	EnergyConversion	BM	7737	7659	1,01	78	194,561		
	EnergyConversion	AM	7734	7700	0,44	34	174,471		
5.	FieldEnergy	BM	7751	7681	0,9	70	317,148	56,1	46,4
	FieldEnergy	AM	7743	7711	0,41	32	259,266	66,9	35,6
6.	SpaceShuttlePayloads	BM	16486	16407	0,48	79	217,218	66,7	39,1
	SpaceShuttlePayloads	AM	16487	16437	0,3	50	175,899	74	36,1
7.	TED_Energy	AM	2030	2008	1,08	22	245,851	79,8	23,7
	TED_Energy	BM	2030	2001	1,43	29	300,784	71,5	32,1
8.	TED_Energy2	AM	600	594	1	6	228,465	86,5	17,1
	TED_Energy2	BM	600	594	1	6	289,510	78,5	24,7
9.	TED_Energy3	AM	4684	4664	0,43	20	214,428	64,2	37,7
	TED_Energy3	BM	4684	4666	0,38	18	274,056	52,9	48,6
10.	TED_Energy4	AM	974	969	0,51	5	277,442	71,2	37,5
	TED_Energy4	BM	974	964	1,03	10	300,300	60,1	46,9
11.	TED_Energy5	AM	2928	2916	0,41	12	226,640	64,1	39,4
	TED_Energy5	BM	2928	2899	0,99	29	246,643	55	48,4
12.	TED_Energy6	AM	4826	4814	0,25	12	163,654	72,6	33,1
	TED_Energy6	BM	4826	4795	0,64	31	202,976	62,6	42,9
13.	TED_Energy7	AM	2666	2658	0,3	8	208,919	69,5	34,6
	TED_Energy7	BM	2666	2648	0,68	18	276,808	59,1	45,1